MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21b/2016/07.06.18.43-TDI

EXTRACTING BEHAVIORAL PROFILES FROM CITIZEN SCIENCE USAGE LOGS

Alessandra Marli Maria Morais

Master's Dissertation for the Graduate Program in Applied Computing, advised by Dr. Rafael Duarte Coelho dos Santos, approved in June 29, 2016.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34P/3M32MFH>

INPE
São José dos Campos
2016

PUBLISHED BY:

Instituto Nacional de Pesquisas Espaciais - INPE
Gabinete do Diretor (GB)
Serviço de Informação e Documentação (SID)
Caixa Postal 515 - CEP 12.245-970
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/6921
Fax: (012) 3208-6919
E-mail: pubtc@inpe.br

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21b/2016/07.06.18.43-TDI

EXTRACTING BEHAVIORAL PROFILES FROM CITIZEN SCIENCE USAGE LOGS

Alessandra Marli Maria Morais

Master's Dissertation for the Graduate Program in Applied Computing, advised by Dr. Rafael Duarte Coelho dos Santos, approved in June 29, 2016.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34P/3M32MFH>

INPE
São José dos Campos
2016

Cataloging in Publication Data

Morais, Alessandra Marli Maria.

M791e Extracting behavioral profiles from citizen science usage logs / Alessandra Marli Maria Morais. – São José dos Campos : INPE, 2016.

xxii + 104 p. ; (sid.inpe.br/mtc-m21b/2016/07.06.18.43-TDI)

Dissertation (Master in Applied Computing) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2016.

Guiding : Dr. Rafael Duarte Coelho dos Santos.

1. Citizen science. 2. Data science. 3. Cluster analysis. I.Title.

CDU 004.2

Aluno (a): ***Alessandra Marli Maria Morais***

Título: "EXTRACTING BEHAVIORAL PROFILES FROM CITIZEN SCIENCE USAGE LOGS".

Aprovado (a) pela Banca Examinadora em cumprimento ao requisito exigido para obtenção do Título de ***Mestre*** em

Computação Aplicada

Dr. Nandamudi Lankalapalli Vijaykumar

Presidente / INPE / SJCampos - SP

Dr. Rafael Duarte Coelho dos Santos

Orientador(a) / INPE / SJCampos - SP

Dra. Karine Reis Ferreira Gomes

Membro da Banca / INPE / São José dos Campos - SP

Dra. Daniela Leal Musa

Convidado(a) / UNIFESP / São José dos Campos - SP

Este trabalho foi aprovado por:

( ) maioria simples

(X) unanimidade

São José dos Campos, 29 de Junho de 2016

"Humans are good, she knew, at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent."

Carl Sagan
in *"Contact"*, 1985

Dedicated to my Grandfather,
Rubens Borges Cicilia.

ACKNOWLEDGEMENTS

First, I would like to thank my grandfather Rubens Borges Cicilia, who asked me almost every day about how this thesis was going on, for his encouragement, wisdom and kind words during the time I devoted to this thesis. He was my strength to move on. I also would like to thank my grandmother Maria Aparecida Cicilia (in memoriam) to be an example of patience and faith. I could not forget my grandparents Sebastião Morais (in memoriam) and Maria Aparecida Modesto de Morais who are examples of perseverance and uprightness. I also would like to thank my parents Marli Borges Cicilia Morais and Antorio Morais for their support, patience, unconditional love and care. Similarly, I could not forget my sister Adrielen Maira Morais de Arruda (and her husband) and my brother Alessandro Antonio Morais (and his wife and daughter) to understand my (occasional) bad mood and complaints. They always have the gift of making me laugh especially in difficult days.

Next, I need to thank my advisor Dr. Rafael Duarte Coelho dos Santos, for his confidence in giving me this research work. I could not forget to thank him for being always available to discuss anything, including all personal advices which I will carry to all my life. I also have to thank Dr. Nandamudi Lankalapalli Vijaykumar for all the time given during the writing of this work. I could not forget to thank M. Jordan Raddick for providing us the usages logs of Galaxy Zoo project and I also have to thank all the volunteers and other collaborators from the Galaxy Zoo project. Without them I may never have achieved this.

Before finalizing, I need to emphasize that the program would not be the same without the friendly help of fellow colleagues and friends. It is impossible to list all of them and at the same time unfair. My honest apology to all friends which I am not mentioning in this work. However, I must mention three of them: Ana Ercilia Fernandes Camilo, Andreia Hisi and Viny Cesar Pereira. I have to thank them for all support, advices, jokes, drawing my attention (all from Andreia Hisi), opinions and wonderful happy hours (cheers!). They were my first readers (sometimes against their will).

Lastly, but not least, my sincere recognition and gratefulness to the National Institute for Space Research (INPE), for the opportunity provided and the Brazilian Research Council (CNPq), for the financial support.

ABSTRACT

Citizen science projects are those which recruit volunteers to participate as assistants in scientific studies. These projects are a longstanding tradition of volunteers recruitment which predates the Internet. The advent of the Web enabled the citizen science projects to expand into new domains and gain popularity. Web-based citizen science is established on technological and motivational pillars. Understanding the motivational aspect for volunteers is crucial to plan, design and manage citizen science projects. Some researchers have studied volunteers' motivation to work as assistants by conducting interviews with selected subgroups. These studies can elicit detailed information from volunteers, but they are restricted to a subset of participants. Another way to infer some information about the volunteers' motivations consist of analyzing records (of which volunteer did what and when) registered by web-based Citizen Science projects. This work aims to investigate information that can be extracted from these records (usage logs), especially those which may help understanding volunteers' motivation. To achieve it, this work adapts a model for human interaction with technology in a citizen science context. The adapted model allows the definition of a set of features which will be used in an attempt to characterize volunteers' profiles. To conduct this research machine learning algorithms and exploratory data analysis will be used following a data science approach.

Keywords: Citizen Science. Data Science. Cluster Analysis.

EXTRAINDO PERFIS COMPORTAMENTAIS ATRAVÉS DE LOGS DE UTILIZAÇÃO DE CITIZEN SCIENCE

RESUMO

Projetos de ciência cidadã são aqueles que recrutam voluntários para participar como assistentes em estudos científicos. Esses projetos são uma tradição de longa data que antecede a Internet. O advento da *Web* permitiu que os projetos de ciência cidadã expandissem em novos domínios e ganhassem popularidade. A ciência cidadã baseada na Web é estabelecida nos pilares tecnológico e motivacional. Compreender o aspecto motivacional dos voluntários é fundamental para planejar, projetar e gerenciar tais projetos. A motivação dos voluntários para trabalhar como assistentes tem sido estudada através da realização de entrevistas com voluntários. Estes estudos podem extrair informações detalhadas dos voluntários, mas são restritos a um subconjunto de participantes. Uma outra maneira para inferir informações sobre a motivação dos voluntários consiste em analizar registros (do que o voluntário fez e quando) coletados por tais projetos. Este trabalho tem como objetivo investigar as informações que podem ser extraídas a partir desses registros (logs de uso), especialmente aquelas que possam ajudar a compreender a motivação dos voluntários. Para alcançá-lo, este trabalho adapta um modelo da interação humana com tecnologia no contexto da ciência cidadã. O modelo adaptado permite a definição de um conjunto de características que irá ser utilizado na tentativa de caracterizar perfis de voluntários. Para conduzir esta pesquisa algoritmos de aprendizado de máquina e análise exploratória de dados serão utilizados seguindo um processo *Data Science*.

Palavras-chave: Ciência Cidadã. *Data Science*. Clusterização.

LIST OF FIGURES

Page

LIST OF TABLES

Page

CONTENTS

Page

1 INTRODUCTION

Over the last few decades advances in different fields of technology have provided new mechanisms for collection and storage of data (KEIM; KRIEGEL, 1995). Such advances enabled significant increase in data volumes in different sectors of society like industry, commerce and science. Currently, it is common to deal with an overload of data and information.

In science, some projects face the challenge to get information from collected data. According to Gray et al. (2005), scientific data is doubling every year, whereas the number of professional scientists available to interpret the data grows much more slowly (RADDICK et al., 2009). In order to deal with the challenge of data overload, some science teams have delegated aspects of analysis to volunteers - members of the public who participate as assistants in scientific studies. Such approach is traditionally called Citizen Science.

The analysis delegated to volunteers is not complex and only requires their cognitive power, in other words, human abilities related to the mental processes of perception, attention, memory, judgement, reasoning, and visual and spatial processing. Citizen science projects delegate such analysis in forms of tasks. These tasks are performed by volunteers based on specific instructions provided by scientific team. As a result of the analysis, volunteers generate data.

Quality of generated data is one of the main issues in these projects; the science team must be prepared to scrutinize the data carefully to discard suspect or unreliable data (COHN, 2008). Besides data quality, another issue of concern is volunteers' motivation. Understanding the motivational aspect is crucial to the plan, design and the management of citizen science projects (NOV et al., 2011; RADDICK et al., 2013). Once the reason which drives volunteers to help a project is known, efforts can be guided to attract more volunteers and keep them collaborating.

Volunteers' motivation to work as assistants have been studied through surveys and interviews (RADDICK et al., 2010; ROTMAN et al., 2012; RADDICK et al., 2013). Although these studies can elicit detailed information from volunteers, they are restricted to a subset of participants of citizen science projects, namely, those who volunteer to answer the surveys and interviews.

1.1 Motivation

There is a significant amount of web-based citizen science projects, as shown by Zooniverse web site[1] and others. These projects are able to collect data of interaction between volunteers and the tasks management interface. At a minimum, data such as, who (anonymous IDs), what (the volunteer's collaboration) and when (timestamps of registered collaboration) a task was performed by the volunteer. Such data (usage logs) may help to understand volunteers' interaction with these projects.

Understanding interactions may help to deal with some issues and challenges of citizen science projects. It may assist attract more volunteers to the project and keep them engaged, understand the reasons for volunteers' abandonment (and possibly suggesting strategies to prevent this), categorize volunteers according to different parameters with the objective to increase data quality by prioritizing or penalizing some collaborations, provide implicit understanding on motivation, facilitate collaboration with the systems and design better user interfaces, and allow better planning and deployment of similar projects and systems (MORAIS et al., 2013).

Usage logs may yield less detailed results about individuals, and be much more restricted in scope when compared to surveys or direct interviews. On the other hand, it is applicable to all registered volunteers. Moreover, it can be enriched by other sources of information such as blogs and websites kept up by scientific team.

Different investigations were conducted from usage logs (MAO et al., 2013; PONCIANO et al., 2014; PONCIANO; BRASILEIRO, 2015) and the authors seem to agree to the fact that few endeavors have been done to take advantage of such records, mainly about volunteers' behavior. Using supervised learning, (MAO et al., 2013) focused on predicting that a volunteer will abandon the project within a given number of tasks or minutes. Following another approach, (PONCIANO; BRASILEIRO, 2015) proposed four measures to be calculated from usage logs and used them as input to clustering algorithms, aiming to find groups of volunteers with similar behavior.

1.2 Objectives

The main objective of this work consists of investigating useful information that can be extracted from usage logs, especially those which may help volunteers' motivation comprehension.

[1] https://www.zooniverse.org/

To achieve it, this work has the following secondary goals:

- Propose a set of measures calculated from usage logs (features), providing more details about the volunteers;
- Explore machine learning algorithms, focusing on clustering methods, which may help extract volunteers' profiles;
- Assess the existence of behavioral profiles with higher collaboration quality compared to others.

1.3 Organization

This thesis is organized as follows. Chapter 2 presents the concepts of citizen science literature, the challenges and issues of this research field, the related works and the focus of this work. Chapter 3 describes the methodology used to conduct this research work. Chapter 4 presents some visual techniques which may provide useful information about volunteers' behaviors. It also presents some statistics and comments about the features proposed, highlighting the first insights about the data structure. Chapter 5 presents the basic concepts of the cluster analysis literature, defines the experiments conducted and describes the results. Chapter 6 proposes an approach to extract volunteers' profile as an alternative to standard clustering methods and uses it to evaluate the issue of profiles quality of collaboration and the profiles found. Finally, Chapter 7 presents the conclusions, contributions and suggestions for future work.

2 CITIZEN SCIENCE

The term "Citizen Science" refers to projects that recruit volunteers, known as citizen scientists, to participate as assistants in scientific studies (COHN, 2008). Such volunteers are often not paid for their assistance, besides not being necessarily scientists. The assistance is done through creation or annotation of data through execution of tasks. These tasks are based on specific instructions, avoiding subjective aspects like opinion, feeling and creativity, so that, such tasks may be evaluated by a professional scientist in terms of correctness.

Task assigned to volunteers are varied. Some citizen science projects require physical or in situ observation of the environment, others just require Internet access, motivation to collaborate, and free time. The latter typically involves classification or annotation tasks on specially crafted datasets through a Web interface, collecting information that can't be easily obtained without human input, such as classifying different animals caught in millions of camera trap images (Snapshot Serengeti), finding and marking "fans" and "blotches" on the Martian surface (Planet Four), helping the analysis of storms through patterns recognition from satellite images (Cyclone Center) and doing transcription of old digitalized documents (Operation War Diary).

The fact that citizen science projects intentionally place responsibility for creating data into the hands of non-experts may seem antithetical, once data-based scientific research require very high quality data. However, at least three reasons can be pointed out to justify these endeavors. First, volunteers can play an important role in reducing costs associated with research projects (BROOKING; HUNTER, 2011), so that, it is possible to conduct data collection surveys without the need to hire many assistant researchers to do what the volunteers do (DROEGE, 2007). Second, scientific data is currently doubling every year (GRAY et al., 2005), whereas the number of professional scientists available to interpret the data grows much more slowly (RADDICK et al., 2009). So beyond reducing costs, citizen science has traditionally been used when some aspect of the data analysis is beyond the capacity of the core science team (RADDICK et al., 2009) as a way to offer a solution to this problem (DROEGE, 2007). Finally, such projects are not a new concept and they have been remarkably successful in advancing scientific knowledge (BONNEY et al., 2009).

The Christmas Bird Count initiated in the 1900s is often cited as the first citizen science project, though some authors describe approaches in the 1800s (DROEGE, 2007) and even in the 1700s (RADDICK et al., 2009). A good example of success

is the Galaxy Zoo project. For most of the twentieth century, morphological catalogues of galaxies were compiled by individuals or small teams of astronomers, but modern surveys like Sloan Digital Sky Survey (SDSS) containing data from millions of galaxies make this approach impractical (LINTOTT et al., 2011). To demonstrate the catalog's quality performed by volunteers, studies were conducted by comparing them with catalog compiled by teams of astronomers and the results were considered satisfactory (LINTOTT et al., 2008).

This project also exemplify how the use of volunteers may surprise and help scientific studies in an unexpectedly way. Hanny van Arkel, a Dutch primary school teacher, found a strange gaseous blob while using the Galaxy Zoo website. This strange object was focus of studies and later it was identified as a quasar ionization echo. Another example which illustrate the use of volunteers as viable in the scientific research process is the discovery of a pulsar by Einstein@home (NOV et al., 2011).

It is also noteworthy that most citizen science projects strive to help volunteers learn about the process by which scientific investigations are conducted (BONNEY et al., 2009). These endeavors benefit volunteers, researches, and society (RADDICK et al., 2010). Volunteers may have fun and increase their knowledge about the research topic and develop their scientific thinking. Researchers are benefited once their projects contain discerning assistants. For society at large, it can build a closer connection between scientists and the public.

This chapter aims to present the bibliography review of citizen science literature with main focus in the issue of volunteers' motivation. Methodologies of interaction with volunteers is categorized on section 2.1. The main challenges and issues which envolve the use of volunteers as assistents in scientific studies is presented on section 2.2. The state of the art of volunteers' motivation is presented on section 2.3. Finally, section 2.4 highlights the topics explored by this work.

2.1 Typology

Citizen Science projects can be classified according to the degree of involvement required from volunteers as low, medium and high level (SOARES, 2011). The low level is called volunteer computing for some authors. On that level, it is not required that the citizen scientists have any kind of knowledge about the subject of the project, their role is summarized by offering computational resources when their computers are idle. Launched in 1999, the SETI@home project[1] is considered as

[1] http://setiathome.berkeley.edu/

being an important precursor of online citizen science projects. This project asks participants to donate idle time on their computers to analyse radio telescope data in an attempt to discover signals from extraterrestrial civilizations.

The medium level describe approaches which require data creation from observations of images. Such approaches involve interaction with task manager though a Web interface. On this level, volunteers have to be aware of their interactions and should use their cognitive ability. Moreover, knowledge about the subject of the project is encouraged. The Galaxy Zoo project[2] is a good example of this level. Started in June 2007 the project asked volunteers to look at pictures of galaxies and report on their morphological features through a website. Such website was designed to recruit volunteers and collect the data provided by them.

Finally, the high level involves data creation from observations made through tasks related to data collection or monitoring. In addition to being aware and using their cognitive ability, volunteers need to physically visit some specific place. Some authors have named this level by Volunteer Sensing (GOODCHILD, 2007) once volunteers often act as sensors which collect geographic data. An example of this approach is the CoralWatch project[3]. It is an attempt to global monitoring of coral bleaching, providing education about coral reef conservation. A chart was developed and validated by the University of Queensland to standardize changes in coral color, providing a simple way to quantify bleaching and monitor coral health (SIEBECK et al., 2006). The project asked volunteers to dive and compare the Coral Health Chart with the observed coral and send their observations through the project website.

2.2 Challenges and Issues

Some issues can be singled out as reasons for the scientific community to perceive citizen science data as low quality and not worthy of being considered in serious scientific research (BROOKING; HUNTER, 2011). The likely source of poor quality, misleading or even suspect data may be summarized by the limited training, lack knowledge and expertise of volunteers, anonymity and lack of commitment.

Data quality is likely the main challenge of a successful project (RIESCH; POTTER, 2013). Mechanisms to enhance the quality and trust of citizen science data have been proposed in literature (SOARES, 2011; ALABRI; HUNTER, 2010; WIGGINS et al., 2011). Nonetheless, some studies have concluded that data produced by volunteers

[2]http://zoo1.galaxyzoo.org/
[3]http://www.coralwatch.org/web/guest

are as good as data produced by professional scientist (COHN, 2008; LINTOTT et al., 2008; BONNEY et al., 2009).

Other challenges in citizen science projects are find ways to attract and retain volunteers with the project until their objects were concluded. These two issues are described as motivational pillar of a citizen science endeavor (NOV et al., 2011). Online citizen science is based on technological and motivational pillars; understanding the motivational aspect is crucial to plan, design and manage citizen science projects (NOV et al., 2011; RADDICK et al., 2013). According Darch (2014) the management of volunteer's behaviors in terms of how they contribute also plays a significant role in improving both the quality of individual contribution and the overall robustness of the resultant dataset.

2.3 State of the Art: Studying Volunteers' Motivation

In pre-Internet citizen science projects, volunteers were recruited to collect data through the observation of the natural world; their observations were often reported via paper forms sent by mail. Such volunteers were usually hobbyists and people who love the outdoors. The advent of the internet enabled volunteers to participate in a new way and projects like SETI@home were launched. Advances in communication technology, specifically the advent of the Web, enabled the citizen science approaches to become even more distributed, expand into new and innovative domains and gain popularity (EVELEIGH et al., 2014). Through Web-based citizen science projects, citizen science endeavors were able to attract and maintain volunteers in any part of the world (NEWMAN et al., 2012), just requiring Internet access, motivation to collaborate, and free time.

The technological aspect of online citizen science endeavors has been extensively studied, whereas the motivational aspect have received relatively little attention if compared to the former (NOV et al., 2011). Motivational aspects seems to have been predominantly analyzed through surveys and interviews with selected subgroups of volunteers. Some examples of such works are Raddick et al. (2013), Reed et al. (2013) and Eveleigh et al. (2014). Although these studies can elicit detailed information from volunteers, they are restricted to a subset of participants in citizen science projects, those who volunteer to answer the surveys and interviews.

Once set in a computational environment web-based citizen science projects are able to store not only data produced by volunteer's collaboration (what), but also the anonymized volunteers (who) and the timestamps of registered collaboration (when).

This kind of information (usage logs), may hold the lifecycle of a citizen science project. Results of policies to get volunteers, actions to keep them collaborating and reason for volunteers' abandonment may be hidden on these records. Therefore, such records may be a powerful tool for feedback to scientific team.

Recent studies have been analyzing such records in attempt to infer information about the volunteers' motivations. It focuses on detection of groups of volunteers whose interactions with the project follow a similar behavioral pattern. This approach might yield less detailed results about individuals and could be restricted in scope compared to surveys or direct interviews, but it's applicable to all registered volunteers. Examples of these works are Morais et al. (2013), Ponciano et al. (2014) and Ponciano e Brasileiro (2015).

2.3.1 Exploring Usage Logs

Although usage logs may hold important information about a citizen science project, few work seem to take advantage of it. In the context of detecting aspects of volunteers' behaviors, Mao et al. (2013) claims that little analytical work has been done on the engagement and disengagement of volunteers. In order to explore the challenge of learning from data, Mao et al. (2013) present studies to predict signs of the attention and effort invested by volunteers. In such study statistical models were designed and constructed to provide predictions about the forthcoming engagement of volunteers.

Following another approach, Morais et al. (2013) and Ponciano et al. (2014), Ponciano e Brasileiro (2015) extract measures from usage logs with the goal of getting a dataset where, for example, each row represents the features of a volunteer. Such dataset is used as the input to machine learning algorithms in an attempt to find volunteers who exhibit similar behavior. The approach adopted by Morais et al. and Ponciano et al. differ on the methodology to find group of similar volunteers.

In an attempt to visualize profiles, Morais et al. (2013) propose a combination of icon-based visualization technique backed up by a Self-Organizing Map (SOM). Such method was applied on log data from the Galaxy Zoo project first release. It allowed characterization of volunteers in different profiles using seven features calculated from measures extracted from the usage logs. Through this approach three profiles could be easily identified: curious (volunteers who joined the project and did most of the collaboration in one or few days, abandoning the project shortly afterwards), curious with potential to become a regular volunteer (volunteers who

had relatively long collaboration period with long intervals between collaboration) and dedicated (volunteers who were assiduous with collaborations well distributed along the collaboration period). The proposed measures were:

- Participation range in days (p_{u^i}): the number of days counted from the first day volunteer u^i interacted with the website until the last recorded interaction of u^i;
- Participation count in days (d_{u^i}): the number of days for which there was recorded interactions of volunteer i with the website;
- Maximum collaboration ($maximum_{u^i}$): the maximum number of collaboration done in one single day by volunteer i;
- Total collaborations ($total_{u^i}$): the number of classifications done by volunteer i during the collaboration period;
- The average of collaborations of volunteers (avg).

In addition to this method, Morais et al. (2013) also proposed a variation of a xy-plot to display the whole data log (nearly 150,000 volunteers and more than 70 million collaborations). That visual technique helped to point out interesting behaviors from group of volunteers as the distribution of attraction and abandonment, the reaction of volunteers for changes in the system, etc.

Ponciano et al. (2014) present a study using usage logs from Galaxy Zoo Hubble and The Milky Way Project. The study focuses on four measures: frequency (number of days which volunteer collaborated at least once), daily productivity (average number of collaboration per day), typical session duration (average period of devoted time to collaboration), and devotion time (total period of devoted time to collaboration). In their work, the authors emphasize that volunteers can be divided into transient (who execute tasks only one day) and regular volunteers (those who return at least one more day). In addiction, such study shows through some plots of their metrics that regular volunteers show a large variation among themselves in terms of metrics. An overview of volunteers' characteristics is also presented.

In a second study, Ponciano e Brasileiro (2015) computed a set of features from the measures described as:

- The number of days elapsed between the day in which the volunteer i joined the project and the day in which the project is concluded (w_i);

- Sequence of dates in which the volunteer i is active (A_i). A volunteer is said active in a day, when he or she collaborate at least once in such date;
- Multiset D_i (an unordered collection of items that may contain duplicates) of the amount of time the volunteer i devotes to the system on each active day;
- Multiset B_i composed by the time elapsed between each two active days. In other words, a multiset of the number of days taken by the volunteer to return to the system to perform more collaboration after an active day.

The authors conducted an analysis using a classical partitioning method (k-means) in an attempt to describe volunteers' profiles. According to them, the results showed that volunteers exhibit five distinct commitment profiles: hardworking, spasmodic, persistent, lasting and moderate. In their study the authors differentiated *helping activity* (an occasional participation) from *voluntarism* (a planned behavior) and figured out features only for volunteers who have been collaborated at least for two different days.

Table 2.1 shows the features proposed by Morais et al. (2013) and Ponciano e Brasileiro (2015) preserving the nomenclature of measures proposed by the authors, some of which are conceptually equivalent albeit calculated differently.

Table 2.1 - Features proposed by Morais et al. (2013) and Ponciano e Brasileiro (2015).

	Morais et. al.	Ponciano et. al.
Relative Activity Duration	$\frac{p_{u^i}}{600}$	$\frac{(Max(A_i)-Min(A_i))+1}{w_i}$
Assiduity	$\frac{d_{u^i}}{p_{u^i}}$	$\frac{mod(A_i)}{(Max(A_i)-Min(A_i))+1}$
Distribution of Collaboration	$\frac{maximum_{u^i}}{total_{u^i}}$	—
Measure of Collaborations	$\frac{total_{u^i}}{avg_{u^i}}$	—
Activity Days	d_{u^i}	—
Total of Collaborations	$log_{10}(total_{u^i})$	—
Daily devoted time	—	$avg(D_i)$
Variation in periodicity	—	$sd(B_i)$

2.4 Scope of this Work

According to Eveleigh et al. (2014), researchers in citizen science assumed that sustained contribution by individual volunteers is critical for the viability of these projects. Therefore, even though researchers are aware that volunteers often slow down or drop out after an initial flurry of activity, they still encourage committed involvement rather than facilitating occasional collaboration. Following a qualitative research through interviews and surveys, Eveleigh et al. (2014) reveals that occasional volunteers might be less motivated compared to super contributors (those who join the project for various days), but they are still motivated and they care about the progress of the project and the quality of their work. Hence, inspired by the conclusions of Eveleigh et al. (2014), this work aims to assess the existence of correlation between quality collaboration and volunteers' behavior.

In the context of guaranteeing the quality of collected data, works like Arcanjo et al. (2014) and Arcanjo et al. (2016) have faced the issue of evaluate volunteers' collaboration. Such works have characterized volunteers through basic statistics about the amount of collaboration and their hit with the right answer. Once the volunteers are characterized, this information is used to weigh their collaborations and motivate the best volunteers by ranking the top volunteers. However, as far as we know, to date there is not a study which assess the existence of behavioral profiles with higher collaboration quality compared to others.

Building on the experience obtained during the research for this dissertation, we propose that volunteers' behaviors can be described through a set of features that can provide more details about how and why volunteer interact with web-based citizen science projects. In order to define that set of features, this work combines the concepts of interaction with technology proposed by O'Brien e Toms (2008) with the suggestion of how a work session can be defined (MAO et al., 2013). Details about this approach will be described on Chapter 3. Increasing the amount of features may bring challenges on the identification of groups of volunteers with similar behavior, so different methods of grouping data (with analysis of their strengths and weaknesses) are studied to achieve a satisfactory approach.

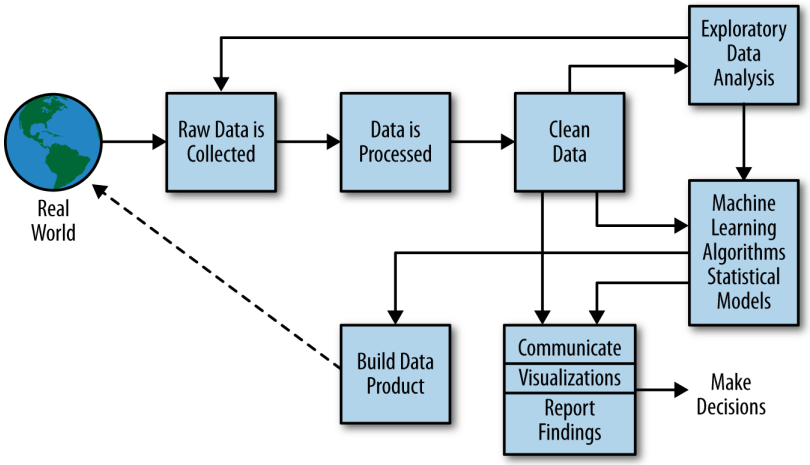
3 DATA AND METHODS

The main objective of this work consists of investigating useful information that can be extracted from usage logs (i.e. records) collected by web-based citizen science projects. As secondary goals this work aims to provide a set of features computed from these records which enable the understanding of the volunteers' behaviors, and to explore techniques which may help extract volunteers' profiles and assess the correlation between quality collaboration and some volunteers' profiles.

3.1 Data Science

To conduct this research a data science process is used as methodology. The general data science framework is shown in Figure 3.1 (O'NEIL; SCHUTT, 2013).

Figure 3.1 - The data science process.



SOURCE: O'Neil e Schutt (2013).

In the context of this work, the "Real World" component of that framework corresponds to people (potential assistants) and science teams developing web-based citizen science projects and mechanisms to attract, involve and interact with their assistants through forum, blogs and social media. Some people decide to collaborate with citizen science projects and become assistants by curiosity, fun or a range of other factors. In a data science process the real world provides raw data from different sources and types. With respect to the scenario of web-based citizen science

projects examples of data sources are: logs, posts in forums, news and web sites. The process starts by selecting the raw data which will help answer the question considered in this research: "*How volunteers interact with a web-based citizen science project?*".

The following sections describe each stage (shown in Figure 3.1) that compose a data science approach and the corresponding activities which will be done to achieve the goals this work.

3.2 Raw Data

In order to assess some techniques which may be useful to answer the question of how volunteers interact with a web-based citizen science project, we will use the usage logs of the Galaxy Zoo first release as a case study. Started in June 2007, the Galaxy Zoo project is a web-citizen science project which presents images from galaxies collected from Sloan Digital Sky Survey (SDSS). Volunteers had to look at pictures of galaxies and report on their morphological features. The project allowed the classification of nearly one million galaxies until 2008 (LINTOTT et al., 2008) and has been considered a success case for citizen science applications.

The raw data recorded by the Galaxy Zoo project contains the records of which volunteer (not anonymized) did what and when, the IP address and the browser used by the volunteer. The raw data available to this work contains the registers of which volunteer (anonymized) did what and when. Such data covers the period from Galaxy Zoo launch in July 8th, 2007 until July 7th, 2012 and holds 79,265,697 collaborations done by 146,669 volunteers.

3.3 Data Cleaning and Preprocessing

Before starting any analysis raw data must be processed to make it clean for analysis. In data science it means that a pipeline should be built and used to scrap, filter and format the raw data, making it more appropriate for further processing (O'NEIL; SCHUTT, 2013). This work adopts the following processing steps:

a) Remove bogus collaborations: Automated mechanisms or some unknown problem with volunteer's browsers may send multiple classifications of a given galaxy for the data collection server, resulting in duplicate entries on usage logs. These classifications were observed by Lintott et al. (2008). In this work, we preserved the first collaboration registered and removed all

others. Details about the process to remove these collaborations and the impact on raw data are described in Chapter 4, section 4.2.

b) Extract measures of interaction: On this step the usage logs are processed to create a dataset where each row describes the features (measures of interaction) of one volunteer. Section 3.3.1 presents the concepts used to extract the set of feature proposed by this work.

3.3.1 Measuring Interaction

O'Brien e Toms (2008) describe the interaction with technology as a process comprised of four distinct stages: point of engagement, period of engagement, disengagement and re-engagement. The authors studied only some aspects of technology, namely, online web-based systems (games, educational sites, shopping, searching), which we consider can be adapted to web-based citizen science tools or portals.

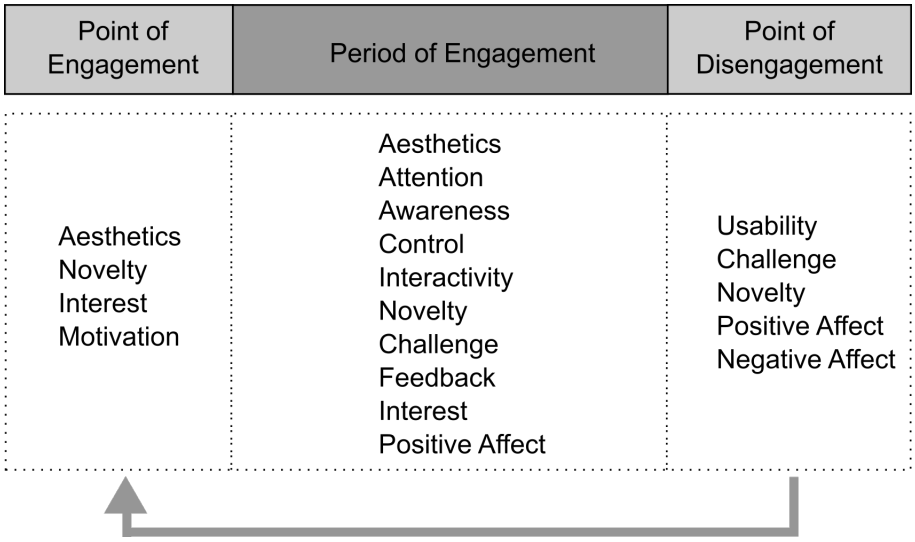
The term engagement is defined by the authors as a quality of user experiences with technology that is characterized by the engagement attributes. An engagement attribute is a characteristic that influences the user experiences or is a component of it. The intensity of these attributes may vary according to users' expectation and the technology itself. Figure 3.2 shows the four stages and the engagement attributes inherent at each stage in that process.

The interaction process starts on the point of engagement. On this stage, something captures the users' attention and moves them forward into a period of engagement. It can be the visual beauty or attractiveness of the computer-based environments (aesthetics), a novelty factor, or something that resonate with the interests of user like learning more about a subject or the satisfaction of achieving a specific goal (motivation).

The period of engagement is marked by the continued maintenance of the users' attention and interest in the interaction. It is achieved through visual components (aesthetics) that keep attention and interest, positive effect as enjoyment and fun, control of interactivity, feedback from the application, novelty factor, and some level of challenge with encourage interaction. On this stage, some participants may not realize that time passed very quickly and they may lack the awareness of physical surroundings.

Once in a period of engagement, disengagement occurred through an internal decision to stop the activity or by interruption and distractions in the physical environ-

Figure 3.2 - Engagement model and its attributes.



SOURCE: O'Brien e Toms (2008).

ment. As reasons to stop the activity, O'Brien e Toms (2008) point out: inability to interact with features of the technology or manipulate interface features (usability), lack or too much challenge, lack of novelty in the application, negative effect (i.e., frustration with technology, boredom and information overload), feelings of success and/or duty accomplished (positive effect), not having sufficient time to interact with or time to devote to the application.

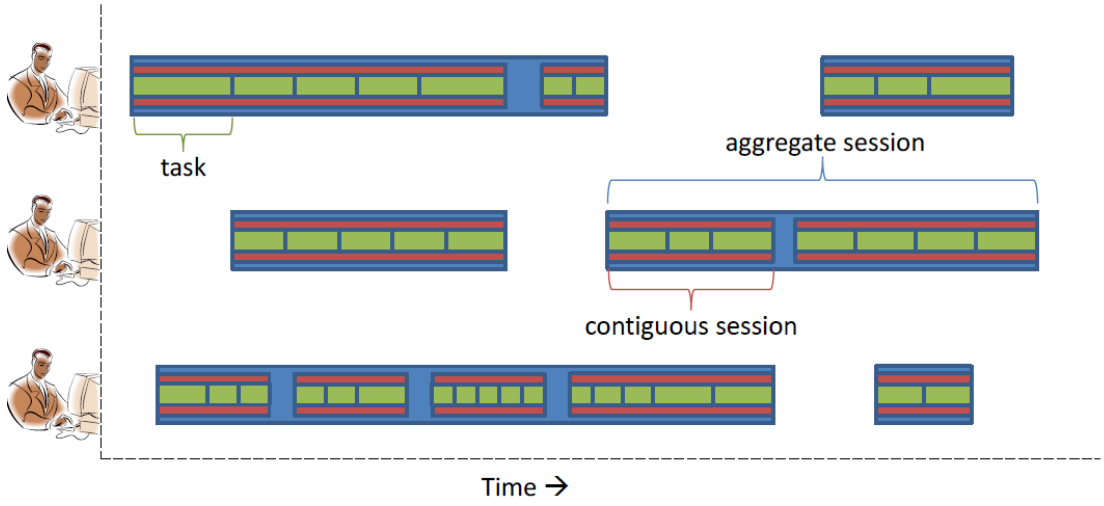
A disengagement decision do not necessarily mean that a user will not start the interaction process again. Users can decide to re-engage in short-term and in long-term. Short-term reengagement is typified by the user who abandoned their tasks as a result of personal needs, or because the user needs time to consider or compare information. Return to an application in the future (long-term) is caused by combination of positive past experience with the application and engagement attributes which compose the point of engagement.

Based on the experience obtained during the research for this dissertation, we claim that such process can shed light on what influences the volunteers' behavior when using web-based citizen science applications and guides the extraction of features from usage logs. To achieve it, we rewrite this process following the definition of work sessions suggested by Mao et al. (2013). In that study the authors consider ap-

plications of crowdsourcing – a term more general than citizen science that includes paid work, and that can be defined as an online, distributed problem-solving and production model that leverages the collective intelligence of online communities to serve specific organizational goals (BRABHAM, 2013).

Mao et al. (2013) defines work sessions though the concept of task, contiguous session and aggregate sessions as shown in Figure 3.3. A task is defined as the smallest indivisible unit of work that can be completed, namely, a single collaboration in a citizen science approach. Sessions of a volunteer consist of the periods of time that a volunteer is engaged on the online platform. Short breaks, where a volunteer intends to return, bring together a sequence of contiguous tasks into contiguous sessions of uninterrupted work. Volunteers can also decide to stop collaborating for longer periods of time; these longer pauses divide the activity into aggregate sessions, comprised of one or more contiguous sessions.

Figure 3.3 - Model of worker sessions.



SOURCE: Mao et al. (2013).

Using data from Galaxy Zoo, Mao et al. (2013) defined the end of a contiguous session as a break of more than five minutes, since it is unlikely for a volunteer to spend this amount of time on a Galaxy Zoo task without activity, namely, analysing the current task without deciding to take an action. Aggregate sessions are consid-

ered as being composed by a sequence of contiguous sessions with intervals between the end of the previous contiguous session and the beginning of the next contiguous session shorter than thirty minutes. It should be noted that based on the experience obtained during the study of the examples of web-based citizen science approaches, such concepts seem be applicable in a wide range of citizen science endeavors.

According to Mao et al. (2013) contiguous and aggregate sessions may have different properties in terms of the engagement of a volunteer. Volunteers are likely to maintain the cognitive context of previous tasks for contiguous sessions that start soon after the end of the prior session; a new session started after the end of aggregate session can be assumed as a new mental context. In other words, the longer the period of time a volunteer spend without collaborating, the less likely this volunteer maintains his cognitive context.

In order to combine the model of O'Brien e Toms (2008) with the concepts defined by Mao et al. (2013) and to avoid the different properties in terms of cognitive context, this work describes the model of engagement in a web-based citizen science context as following:

a) **Point of engagement:** correspond to the first collaboration which may move a volunteer forward into a period of engagement;

b) **Period of engagement:** characterized by a contiguous session, namely, a sequence of contiguous tasks interrupted by short breaks (intervals shorter than five minutes) where workers intend to return to the task;

c) **Point of disengagement:** characterized by the lack of activity for longer than five minutes, since five minutes without activity characterize the end of a contiguous session;

d) **Re-engagement:** new collaborations made after more than five minutes after the point of disengagement. The interval for a re-engagement can be labeled as short-term (greater than five and less than thirty minutes), medium-term (greater than thirty minutes and less than twenty four hours) or long-term (greater than twenty four hours).

Most of engagement's attributes described by O'Brien e Toms (2008) cannot be measured through usage logs. Attributes like interest, motivation, positive or negative affect, and perceived time are subjective and pertain to volunteer's feeling;

other as change on aesthetic, insertion of new tasks and lack of feedback may not be measured through usage logs. On the other hand, logs allow the measure of other interesting and potentially useful features, described in the remainder of this section.

A set of measures which may be easily calculated from usage logs, for each volunteer, is: participation range in days (P), calculated as the number of days counted from the first day the volunteer started collaborating with the project until his/her last recorded interactions; participation count in days (D), namely, the number of days for which there was recorded interaction for a particular volunteer; total of collaborations performed by volunteer (C); maximum collaboration in a single day (M); time spent in a period of engagement (Δt_{prdX}); and total time devoted, calculated as

$$Ttd = \sum_{x=1}^{|T_{prd}|} \Delta t_{prd}x, \text{ where } T_{prd} = \{\Delta t_{prd1}, \Delta t_{prd2}, ..., \Delta t_{prdN}\}.$$

Based on these measures, the following features may be described:

a) ***Relative Activity Duration***: is calculated as the ratio $\frac{P}{w}$ (MORAIS et al., 2013; PONCIANO et al., 2014), with values within $(0, 1]$, where w is the amount of days observed from the first collaboration. Values close to zero characterize volunteers who joined and left the project soon thereafter, while values near to one indicate volunteers who contributed for a long time;

b) ***Assiduity***: also called *Activity Rate* (PONCIANO et al., 2014), it is calculated as $\frac{D}{P}$, with values within $(0, 1]$ and, can be used as a simple frequency of collaboration feature. Values near to one indicates that the volunteer collaborate every day with the project;

c) ***Distribution of Collaboration***: calculated as $\frac{M}{C}$ (MORAIS et al., 2013). It is a rough measure of regularity or pace of activity. Values near to one indicate volunteers who did almost all classifications in a single day;

d) ***Maximum Sustained Effort***: calculated as $max\,(T_{prd})$, it measures the major interval of time in that a volunteer maintains his/her mental context collaborating. This feature may indicate intense periods of dedication to the project and, in extreme cases, may indicate possible attempts to automate the process by, e.g., bots;

e) ***Processing Power***: calculated as $\frac{C}{Ttd}$, this feature describes how many tasks a volunteer executes per second;

f) ***Novelty***: this feature is an attempt to represent the novelty engagement attribute, related to the period of engagement. To measure it, we consider that a volunteer will recognize two tasks as being the same if they occur on the same period of engagement. Therefore, given the set $Nv_{prd} = \{Tnv_1, Tnv_2, ..., Tnv_m\}$, where Tnv_x is the amount of distinct task in Xth period of engagement, novelty is calculated as:

$$\frac{\sum_{x=1}^{|Nv_{prd}|} Tnv_x}{C}$$

With values within $(0, 1]$. Values near to zero denote lack of novelty, while values near to one indicate novelty;

g) ***Challenge***: is another engagement attribute related to the period of engagement. To measure challenge this work adopts that a task is challenging if volunteers agreement is low and easy if the opposite. Therefore, given the set $Ch_{prd} = \{Tch_1, Tch_2, ..., Tch_m\}$, where Tch_x is the amount of challenge task in Xth period of engagement. Challenge is calculated as:

$$\frac{\sum_{x=1}^{|Ch_{prd}|} Tch_x}{C}$$

With values within $[0, 1]$. Values near to zero denote lack of challenge, while values near to one indicate much challenge;

h) ***Recurrence***: this feature indicates if a volunteer has more than a period of engagement by active day. Let Tst be the number of re-engagement that occurred during a short-term interval and Tmt be the number of re-engagements that occurred during a medium-term interval, recurrence is calculated as:

$$\frac{Tst + Tmt}{D}$$

i) ***Agreement***: calculated as $\frac{Ag}{C}$, where Ag is the amount of performed tasks which complies with the opinion of the majority of volunteers who performed the same task. Lintott et al. (2008) proposed the concept of agreement as a measure of correctness, but highlighted that agreement of the volunteers with a certain task does not necessarily mean that the achieved result is correct.

Throughout this work, we consider the first eight measures (a to h) as being the features able to abstract the behavioral aspects of volunteers. These eight features will be used during the experiments which seek to uncover groups of volunteers with similar behaviors. Once the different profiles have been found, the feature agreement is used to assess the existence of behavioral profiles with higher collaboration quality compared to others.

3.4 Exploratory Data Analysis

Once the data is clean and processed, exploratory data analysis can be done. John W. Tukey, the definer of the exploratory data analysis (EDA), describes it as "an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there" (BADIE et al., 2011). In EDA the exploratory aspect means that the understanding of the problem is changing while the analysis is happening (O'NEIL; SCHUTT, 2013). The basic tools of EDA are plots, graphs and summary statistics.

On this stage, we aim to demonstrate the kind of useful information that may be hidden on usage logs. Plots, graphs and summary statistics will be used to provide the first insight about the features extracted from usage logs. At the end of this stage, it is expected some expertise about what kind of behavior should be found in the next stage.

3.5 Machine Learning

Machine Learning algorithms are largely used to predict, classify or cluster data. This work intends to make use of clustering algorithms in an attempt to understand the volunteers' behavior. On this stage, different algorithms will be used aiming to find techniques which fits case study and the features proposed.

3.6 Report Findings

At the end of data science process, results are visualized, interpreted and reported.

3.7 Data Product

Data products are what make data science special and distinct from statistics. It can be incorporated back into the real world, and people can interact with that product generating more data, which creates a feedback loop. At the end of this work the main data product is a dataset containing information about the characteristics

of almost 150,000 volunteer: this dataset may be used in other research works to explore other approaches for volunteers' profile extraction. Moreover, the technique proposed in Chapter 6 to help the extraction of volunteers' profiles provides data which abstract the information of almost 150,000 volunteers into a few hundred of data points which may be useful for future investigations involving data from other web-based citizen science projects.

4 LOOKING AT DATA - EXPLORATORY DATA ANALYSIS

This chapter presents the results of exploratory data analysis. Throughout this chapter, data is analyzed in three different levels. Section 4.1 presents two visual techniques which may be applied on the raw data (logs) in order to show general views of the activity of the volunteers. Section 4.2 explains each of the features proposed, with examples, plots and analysis based on Galaxy Zoo project (first release). Finally, section 4.3 describes the first insights on distribution of data composed by these features, with emphasis on finding evidences of groups (clusters) on data.

4.1 An overview of Volunteers' Activities

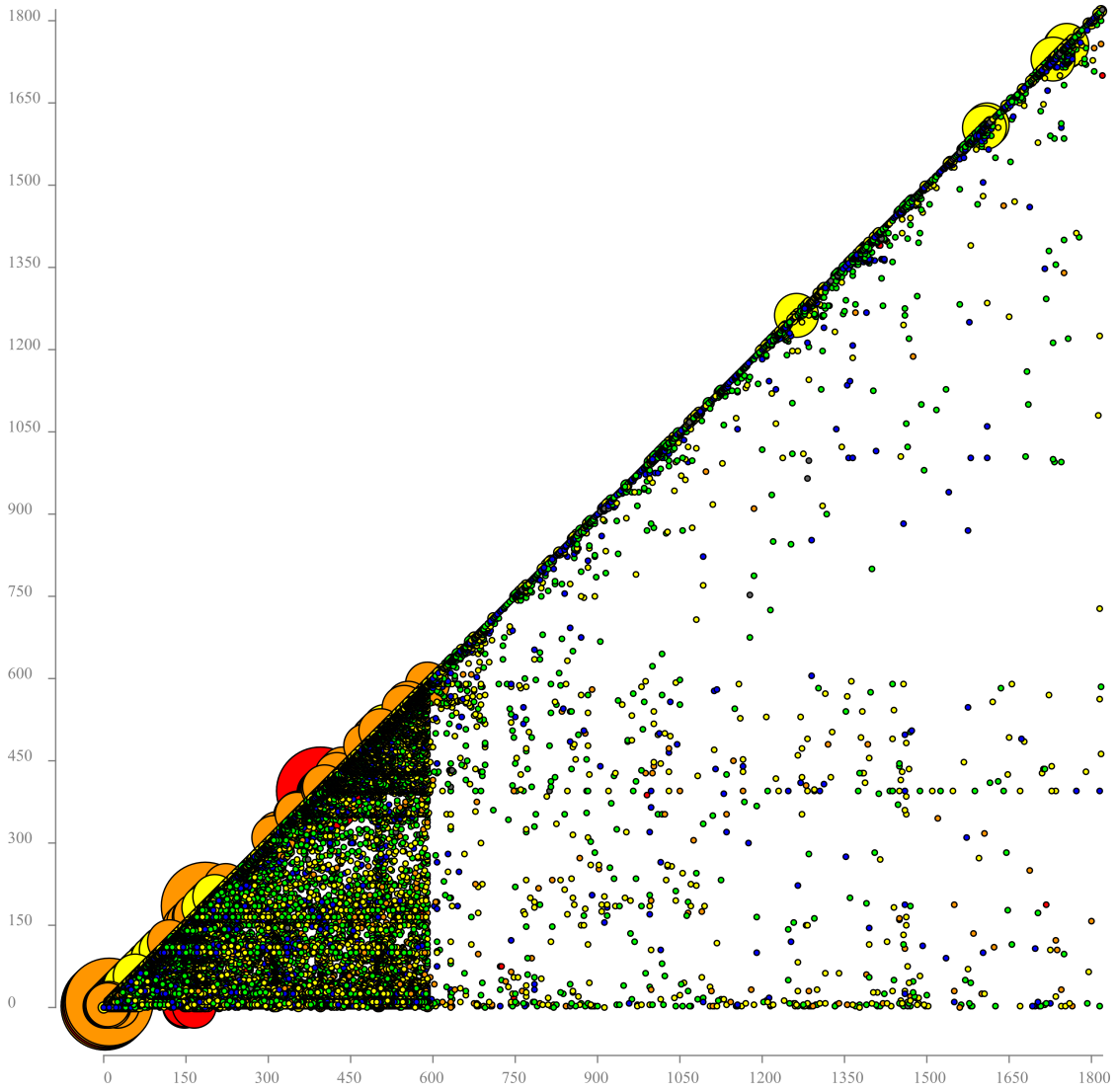
Visualization is usually described as the mapping of data to a graphical representation. Visual representations, if well constructed, can be useful not only to present the information quickly to users, but also to help and maximize data understanding (MAZZA, 2009). In order to get some insights of volunteers' general behaviors during the project's duration, some visualization techniques may be applied on usage logs (MORAIS; SANTOS, 2015). These techniques are useful to highlight information regarding the results of policies to get volunteers, actions to keep volunteers collaborating and volunteers' abandonment.

Figure 4.1 shows an icon-based visualization schema proposed by Morais et al. (2013) to point out interesting behaviors from group of volunteers as, for example, the reaction of volunteers for changes in the system and the distribution of attraction and abandonment. In this schema circles are used as visual icons and are distributed in a XY-plot. Each circle in a XY coordinate represents the group of volunteers that joined (on day Y) and left (on day X) the project, while the radius refers to the number of volunteers in that particular group and the color of the circle corresponds to the total of classifications performed by them. Labels for the radius and colors are shown in Figure 4.2.

The main visible characteristic of Figure 4.1 is the division of the points in two regions with distinct densities, separated around the 600 day's mark; it shows that most people stopped collaborating shortly before day 600. This may be justified due to the possible migration of volunteers from Galaxy Zoo I to a new version of the project (Galaxy Zoo II) announced on February 16th, 2009 (day 589) via the volunteers' mailing list. Other interesting visual characteristics are pointed out in (MORAIS et al., 2013). One of them is the high level of abandonment close to day 150 caused by changing the appearance of images presented to volunteers. Another is the

significant increase on volunteers' access closest to day 400, due to the discovery of an astronomical object of unknown nature whose discovery was reported in several news outlets, including CNN.

Figure 4.1 - Visualization of volunteers' groups size and total collaboration based on their first and last activity in the project. Labels are shown in Figure 4.2.



SOURCE: Morais et al. (2013).

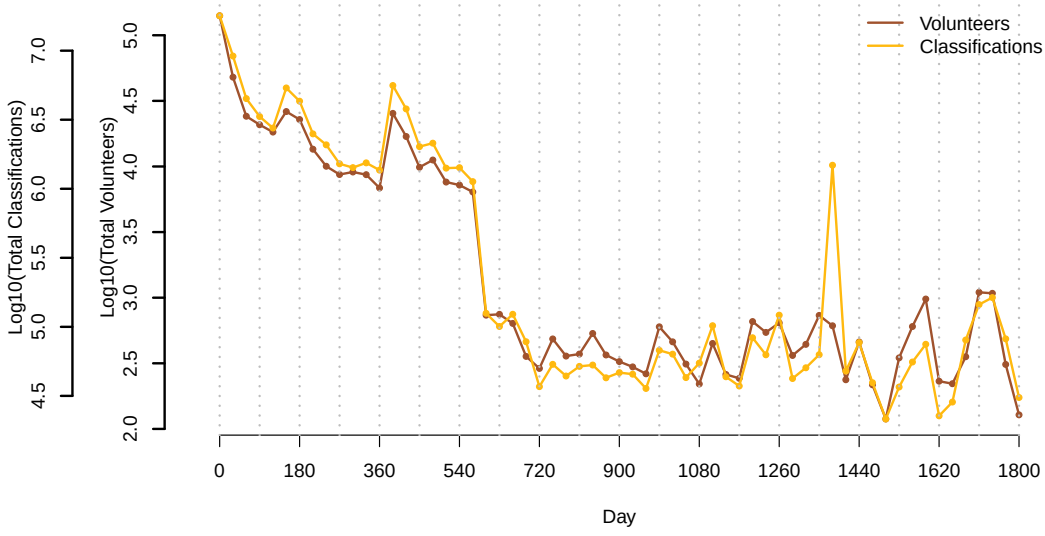
Figure 4.3 shows the evolution of the number of volunteers and classifications during the 1822 days with accumulated values for every 30 days. It is possible to note that for most of the project's duration, the number of volunteers and classifications was more or less proportional. The exceptions and their explanation are presented in

Figure 4.2 - Labels for Figure 4.1

SOURCE: Morais et al. (2013).

detail by Morais e Santos (2015). One of these exceptions are the temporary increase, around day 1,406 of the project (May 13th, 2011), in the number of classifications, but not in the number of volunteers. Figure 4.4 shows this exception with more details.

Figure 4.3 - Number of volunteers and classifications during the 1,822 days covered in the logs. Most of the classifications were done in the first 600 days of the project, with a sharp decline in both the number of volunteers and classifications after that.

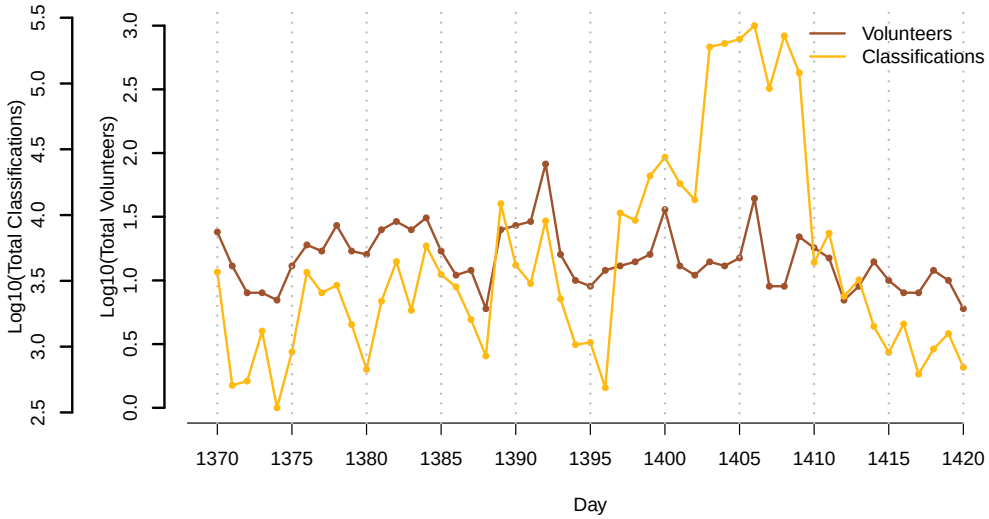


SOURCE: Morais e Santos (2015).

By analyzing the raw data, it was noticed that between days 1397 and 1410, 170

volunteers performed classifications, but only three volunteers were responsible for almost 95% of the classifications during that period. These three volunteers joined the project few days before and left few days after the analyzed period, performing bursts of collaborations during 10 hours or more. The sustained collaboration for long hours indicates a possible attempt to automate classification via a software agent.

Figure 4.4 - Peak classifications around day 1,406.



SOURCE: Morais e Santos (2015).

It is interesting to present some additional information about the data. As pointed out by Morais et al. (2013), 91.78% of the volunteers were active on the data collection website between days 0 and 600, contributing with 90.52% of the total classifications. Of these, 62.82% of the volunteers joined the project and collaborated for just one day, or left after just a few days, making a total of 6,256,407 classifications - 7.8% of the total project ratings. In Figure 4.1 these are shown as circles located over or around the main diagonal line. The 37.18% of remaining volunteers have made a total of 65,525,422 classifications, corresponding to 82.63% of the total.

4.2 Features for Volunteers' Characterization

The first step to calculate the features proposed to volunteers' characterization involves removing false collaborations (unreliable or from unknown sources). These false collaborations may be produced by automated scripts, navigation issues or

other reasons (LINTOTT et al., 2008). Figure 4.5 shows two easily detectable types of false collaborations. In the first type, volunteers performed multiple classifications for a given galaxy always choosing the same option, while in the other case the chosen option seems to be random values, entered in a very short time interval. Analysis of our case study showed that 19,389 (13%) volunteers have at least one sequence of multiple classifications of first type and 10,197 (6.9%) of the second type. We processed the raw data to eliminate these collaborations, removing 477,159 (0.6%) records in this step.

Figure 4.5 - Examples of false collaborations. From left to right separated by semicolon: galaxy id, option chosen and timestamp.

```
587741421632356792;15;2007-10-05 00:52:06
587741421632356792;15;2007-10-05 00:52:06
587741421632356792;15;2007-10-05 00:52:07
587741421632356792;15;2007-10-05 00:52:07
587741421632356792;15;2007-10-05 00:52:08
587741421632356792;15;2007-10-05 00:52:09
587741421632356792;15;2007-10-05 00:52:09
587741421632356792;15;2007-10-05 00:52:11
587741421632356792;15;2007-10-05 00:52:11
587741421632356792;15;2007-10-05 00:52:13
587741421632356792;15;2007-10-05 00:52:14
```

(a) Multiple classifications with the same label

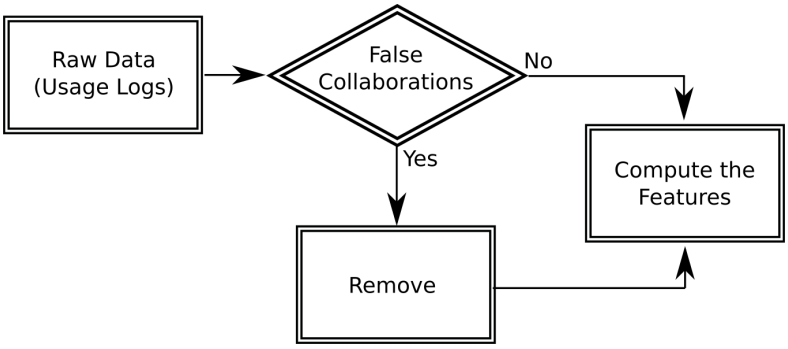
```
587731885732200839;12;2007-10-05 01:02:13
587731885732200839;13;2007-10-05 01:02:13
587731885732200839;14;2007-10-05 01:02:13
587731885732200839;11;2007-10-05 01:02:15
587731885732200839;11;2007-10-05 01:02:15
587731885732200839;11;2007-10-05 01:02:15
587731885732200839;11;2007-10-05 01:02:16
587731885732200839;16;2007-10-05 01:02:16
587731885732200839;16;2007-10-05 01:02:16
587731885732200839;15;2007-10-05 01:02:17
587731885732200839;15;2007-10-05 01:02:17
```

(b) Multiple classifications with random labels

The next step consist of preparing a dataset, where each row describes the features of a given volunteer. Since volunteers may start their collaboration for the first time at different moments of the project, we used an observation window of 600 days to all volunteers. We chose the interval of 600 days because it covers the launch of the Galaxy Zoo I and the announcement of Galaxy Zoo II. From the observation window we selected the records to compute the features proposed in section 3.3.1. Following this approach, volunteers who joined the project for the first time after November 15th, 2010 were discarded because they don't have 600 days to be observed until July 7th, 2012 (the last day covered by the case study) and collaborations which are done after the 600th day (counted from the first active day) of a given volunteer were discarded because they are out of the observation window of this volunteer. In the total 6,952 (4.7%) volunteers and 266,906 (0.33%) collaborations were eliminated. Figure 4.6 summarizes the process to make the raw data appropriate to extract the features.

It should be noted that some of the features proposed depend on the concept of period of engagement. In the analyzed case study around 3% of the total number of volunteers haven't had a period of engagement. Due to this we removed these

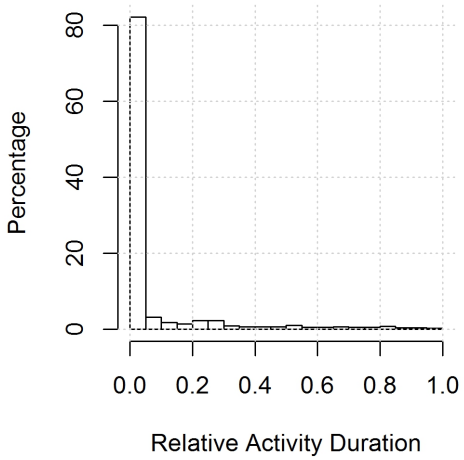
Figure 4.6 - Steps to calculate the features



volunteers. Next we describe each feature to provide an overview of what it may explain about the volunteer's behavior.

Relative Activity Duration: it gives an idea about the life cycle of a volunteer. A volunteer's life cycle is composed by active and inactive days; it starts with the first active day and finish with the last active day. Volunteers who have a short life cycle join and leave the project soon thereafter (values close to zero), while long life cycle is characterized by the volunteers who contributed for a long time (values close to one). Figure 4.7 points out that large number of volunteers has a short life cycle and less than 10% percent were active for more than one month.

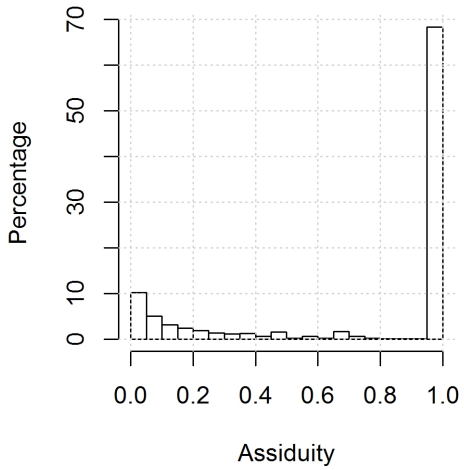
Figure 4.7 - Distribution of Relative Activity Duration



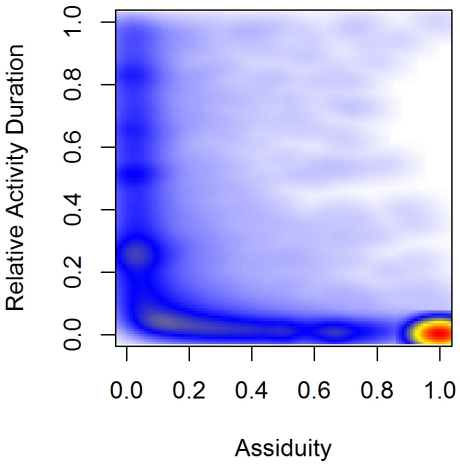
Assiduity: is a feature designed to measure how active a given volunteer was during his/her life cycle. Values close to one indicate that the volunteer collaborates almost

every day with the project, while values close to zero indicate that the volunteer had more inactive days than active days. Figure 4.8(a) shows that almost 70% of volunteers seems to be assiduous, but Figure 4.8(b) shows that assiduity is a characteristic of volunteers who join the project just for fews day (note the red region which consist of the higher density of points). In other words, with the increase of Relative Activity Duration, assiduity seems to be low for the major part of volunteers with some exceptions. One of these exceptions is noteworthy: manual analysis showed that one of the volunteers has a life cycle of 600 days and indicate that this volunteer contributed with the project almost every day (assiduity of 99%).

Figure 4.8 - Assiduity.



(a) Distribution of Values

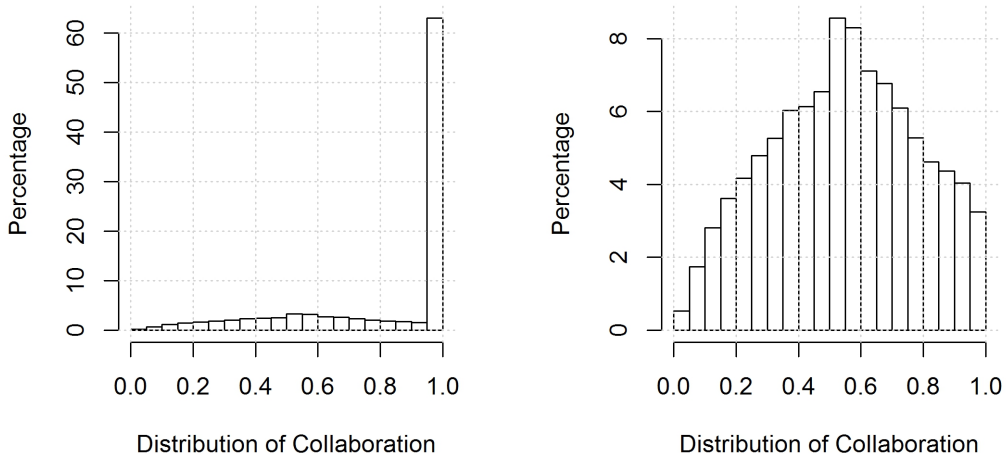


(b) Assiduity relative to Relative Active Duration

Distribution of Collaboration: it was designed to measure how the collaborations are distributed during active days. Values close to zero indicate that the collaborations were well distributed while values close to one indicate that almost all collaborations were done in one day. Figure 4.9(a) shows that more than 60% of volunteers had values close to one. This was expected, since a major part of volunteers joined the project just for one day. Figure 4.9(b) shows the histogram without the volunteers who joined the project just for one day. It highlights that few volunteers spread their collaboration homogeneously on their activity days.

Maximum Sustained Effort: it measures the longest interval of time that a volunteer collaborated with the project while maintaining his/her mental context. Figure 4.10 shows that most of the volunteers (80%) had short periods of sustained

Figure 4.9 - Distribution of Collaboration.

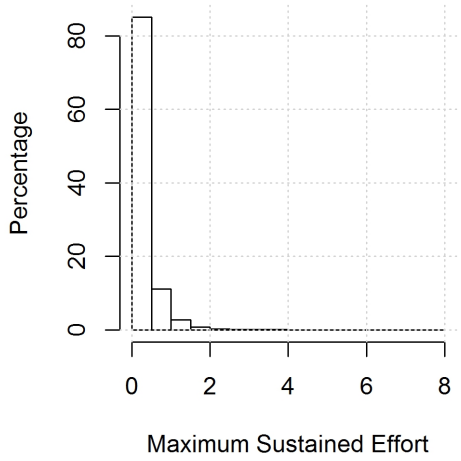


(a) All volunteers

(b) Volunteers who collaborate more than one day

effort (less than thirty minutes) which is consistent with what is expected by human volunteers. However, this feature enables identifying some volunteers with high values of sustained effort. Analysis showed that 0.4% of volunteers spent more than 2 hours of sustained effort, of these, ten volunteers had sustained effort longer than five hours. It indicates possible attempts to automate the process of performing tasks.

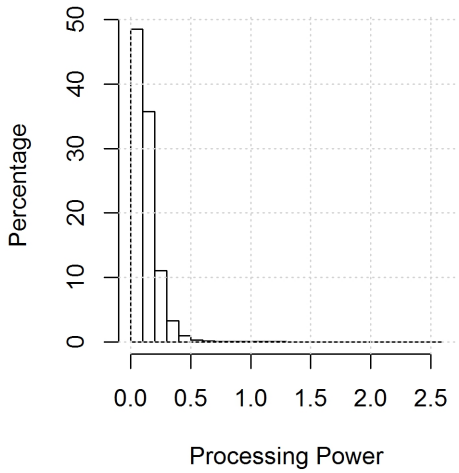
Figure 4.10 - Maximum Sustained Effort



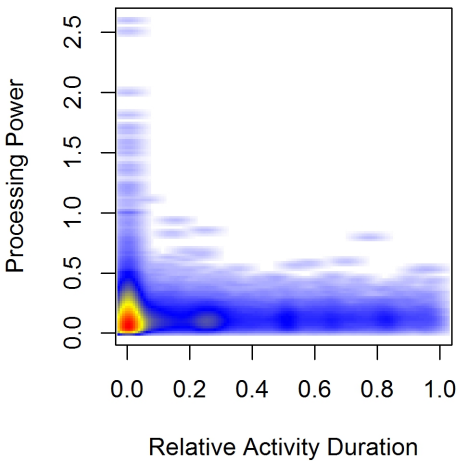
Processing Power: is a rough way to measure the attention attribute proposed by O'Brien e Toms (2008). It measures how many collaborations are done per second.

The histogram in Figure 4.11(a) shows that volunteers usually did few collaborations per second which indicate that they spent, at least, some seconds thinking about the task opting for a choice. However, some exceptions may be clearly observed in Figure 4.11(b). These exceptions occurred with some volunteers who joined the project just for few days. Note that this feature may help to identify either attempts to automatic execution of tasks or volunteers submitting classifications without really paying attention.

Figure 4.11 - Processing Power.



(a) Distribution of values

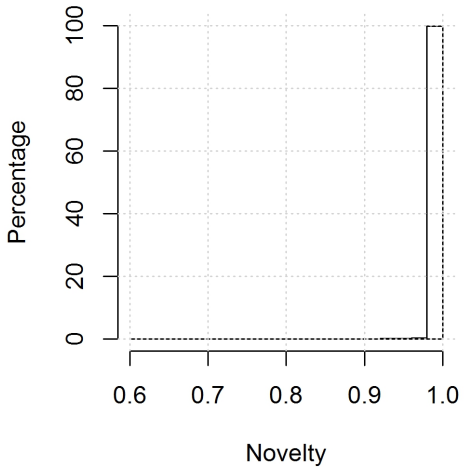


(b) Processing Power relative to the Relative Activity Duration

Novelty: is an engagement's attribute responsible for attract and maintain users interacting with technology. Figure 4.12 shows that the data collection interface usually did not repeat a task during the same period of engagement, since few volunteers have values under 1. It should be noted that if this feature was extracted from the raw data, it would have pointed out low level of novelty for volunteers whose records contain false collaborations. Note that this feature may indicate the necessity of improving the software used to present the tasks to the volunteers.

Challenge: is also an attribute related to the attraction and engagement of volunteers. In the context of a citizen science project, challenge can be inferred based on the agreement of several volunteers on the solution of a given task. An easy task is a task which presents high agreement among the volunteers who performed this task. On the other hand, a challenging task is a task which arouses doubt, resulting in a low agreement among volunteers (considering the Galaxy Zoo project, for example,

Figure 4.12 - Novelty



a task may be considered challenging when 50% of the volunteers say the task refers to an elliptical galaxy and the other 50% say the opposite). Values close to zero denote that the volunteer has to deal with easy tasks and values close to one indicate that his/her tasks were a challenge. Figure 4.13 shows that volunteers usually dealt with easy tasks: note that only few volunteers had 50% of their tasks classified as being difficult.

Figure 4.13 - Challenge

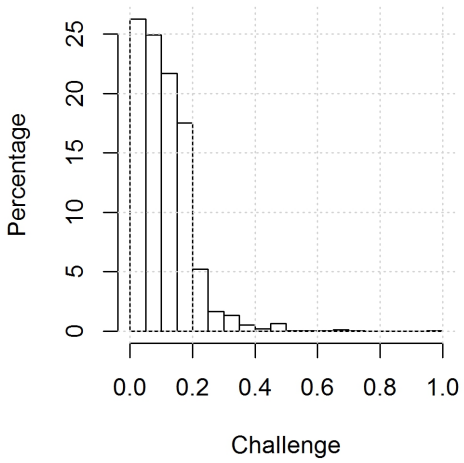
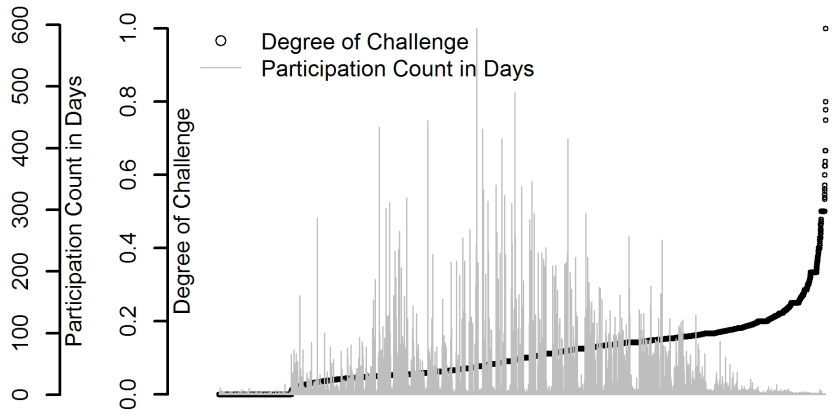


Figure 4.14 shows the values of the challenge feature calculated for each volunteer on the project. On the same figure we plotted the number of days that each volunteer collaborated with the project. Figure 4.14 shows that volunteers who had to deal with very low or very high challenge did not collaborate too long with the project.

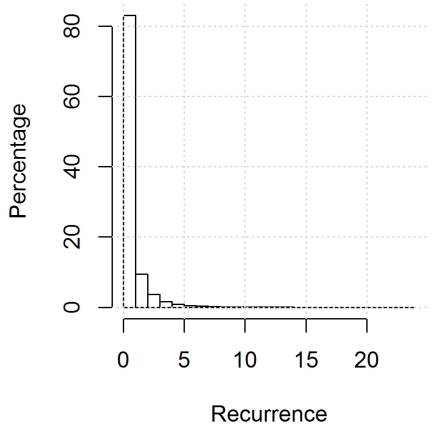
It is not possible to claim that these volunteers abandoned the project early due to the excessive challenge or lack of it.

Figure 4.14 - Challenge and participation count in days



Recurrence: it measures if a volunteer had more than a period of engagement per active day. Zero indicates volunteers who had never returned in the same active day. In the case study, its values range from zero to 24. Figure 4.15 shows that the majority of volunteers had just one period of engagement and very few volunteers had more than 5 periods of engagement in a day.

Figure 4.15 - Recurrence



Agreement: it is an attempt to measure the quality of collaboration of a volunteer. To measure the quality of collaboration with some accuracy, this feature must be

calculated from an auxiliary database obtained by some validation procedure. We use the catalogue proposed by Lintott et al. (2008) to calculate this feature. The catalogue contains the galaxy ID and the feature of the galaxy (spiral, elliptical or uncertain). Based on this catalogue, an answer is said correct if it agrees with the catalogue as being spiral or elliptical. Galaxies labeled as uncertain by the catalogue were not considered to compute this feature. Figure 4.16 shows that more than 40% of volunteers presented a good quality on collaboration (values greater than 0.8).

Figure 4.16 - Agreement.

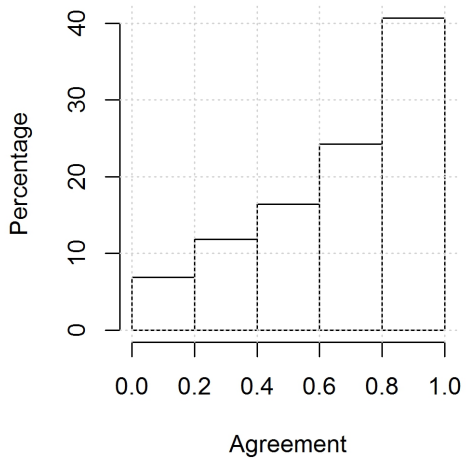
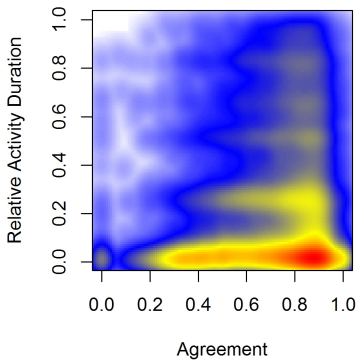
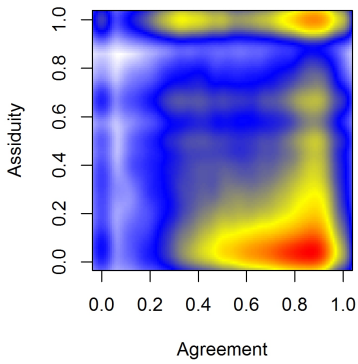


Figure 4.17 shows the distribution of feature agreement in relation to the other features. Through Figure 4.17 we can not observe any characteristic in the measures of interaction that results in a good or a bad quality on collaboration. By observing the Figures 4.17 (a), 4.17 (b) and 4.17 (c), for any value of Relative Active Duration, Assiduity or Distribution of Collaboration there are volunteers with good and bad quality on collaboration. Even if we analyze the regions with high agglomeration of points (in red) we could not point out any relation between the feature in analysis and the quality of collaboration. In Figures 4.17 (d) to 4.17 (h) it is possible to note some outliers, most of them did not point out any evidence of some characteristic which shed light on which results in a good or bad collaboration. Figure 4.17 (d) shows some volunteers with high values of maximum sustained effort (more them six hours) with good quality and some with low quality on collation.

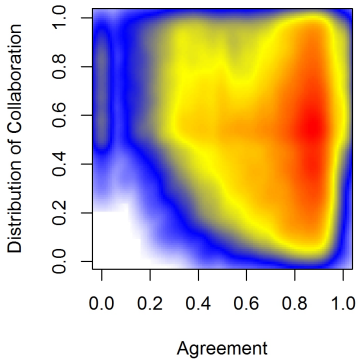
Figure 4.17 - Distribution of feature agreement in relation to the other features.



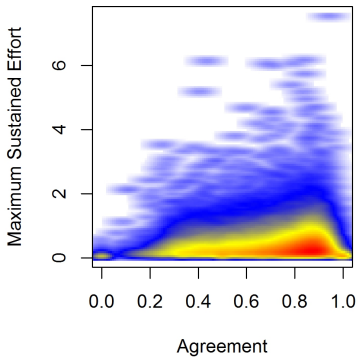
(a) Relative Active Duration



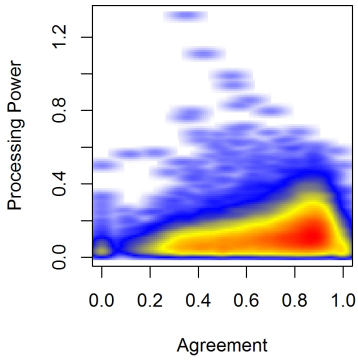
(b) Assiduity



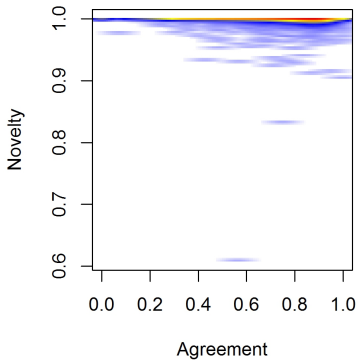
(c) Distribution of Collaboration



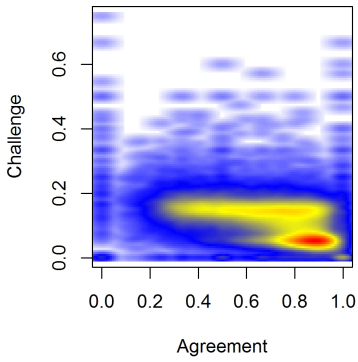
(d) Maximum Sustained Effort



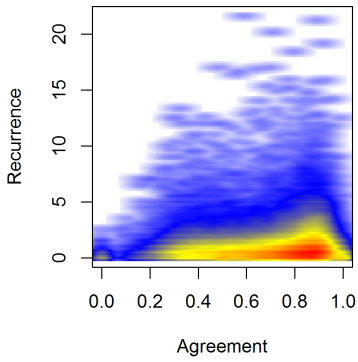
(e) Processing Power



(f) Novelty



(g) Challenge



(h) Recurrence

A possible exception may be observed in Figure 4.17 (e). Note that volunteers who performed more than one collaboration per second have low collaboration, maybe because they did not pay much attention or they were an attempt to automate the process of classifying a galaxy. Volunteer with good collaboration seems to perform low collaborations per second, however, performing few collaborations per second did not ensure good quality; see the points in the left bottom corner.

4.3 Insights on Data Distribution

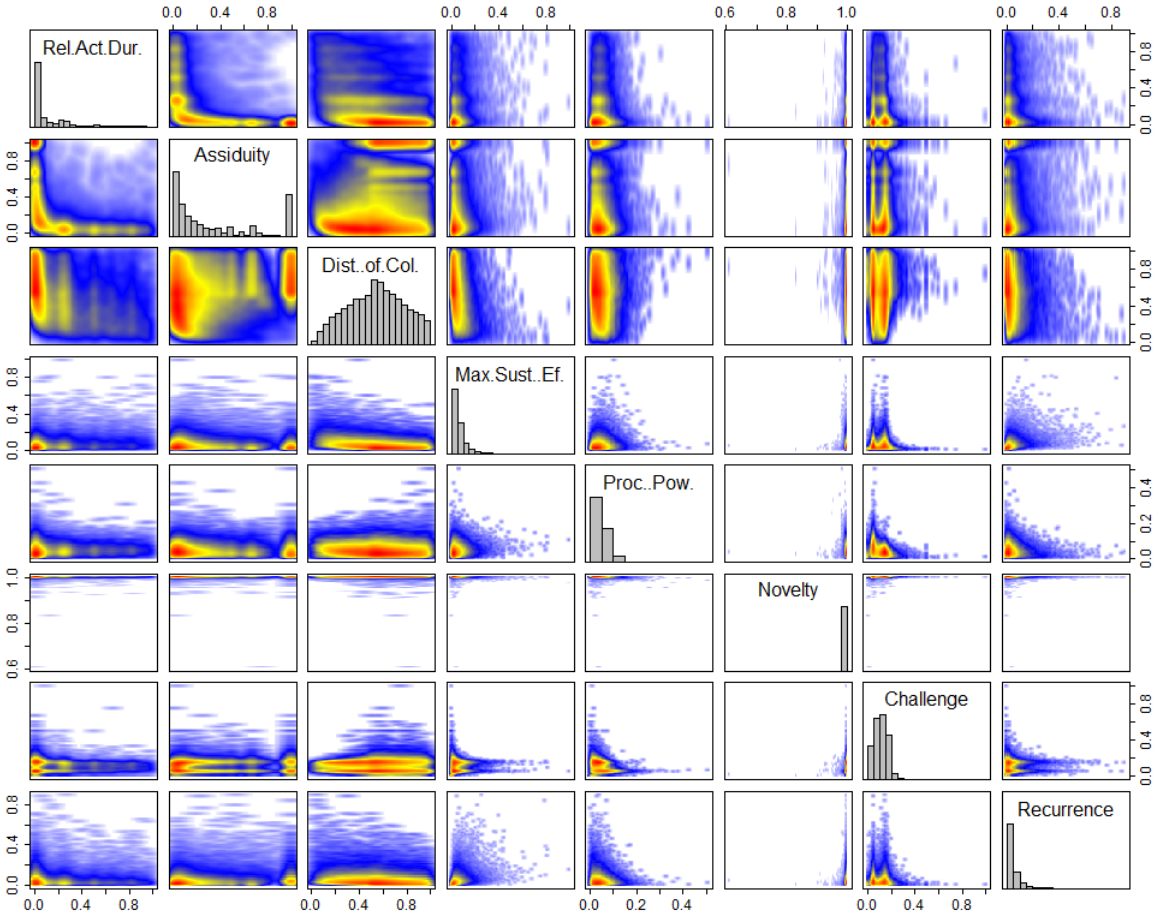
The existence of distinct modes in data distribution is a strong evidence of clusters, while unimodal distributions correspond to homogeneous unclustered data (EVERITT et al., 2011). Data plots are useful tools to provide some insights of the data distribution; it may suggest whether the data contains clusters (modes) and shed light about which clustering methods might be applied for its analysis. There are two kinds of data plots: direct plots and indirect views of data.

Direct plots are obtained using the original variables. Some examples of direct plots used to detect evidence of clusters are histograms, bivariate scatterplots and scatterplots matrices. A scatterplot matrix is defined by Cleveland et al. (1985) as being a symmetric grid of bivariate scatterplots. The grid consists of n rows and n columns, each one related to one of the n variables, where grids' cells show a scatterplot of two variables. According to Everitt et al. (2011), scatterplots matrices are often used as an initial examination of data for informal evidence about some cluster structure.

Figure 4.18 shows a scatterplot matrix with the estimate of data density and histograms in its diagonal. This Figure points out some visual evidence of regions with higher (in red) to medium (in yellow) density of points. In Figure 4.18, data for volunteers that joined the project for a single day were not plotted, so it would be easier to visualize more subtle dense groupings. Figure A.1 on the Appendix shows the same plot with all volunteers' data.

Indirect views can be created by using data from a suitable dimension reduction technique. This approach is used as an attempt to understand the true nature of data structure which may be not reflected by direct two-dimensional views. There are a number of lower-dimensional projections methods; according to Everitt et al. (2011) the most common is Principal Components Analysis (PCA). PCA is a method for projecting the data to a new coordinate system, where the original variables (possibly correlated) are transformed through linear combinations into a new set of uncorrelated variables called Principal Components (PCs). This transformation

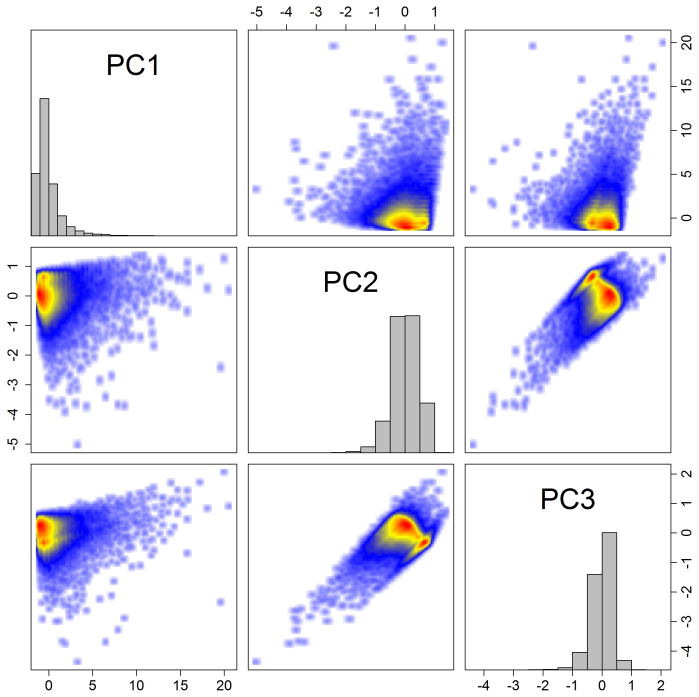
Figure 4.18 - Scatterplot matrix for all the features being considered (excluding data for volunteers who joined the project for just one day). Each cell shows the data density for the combination of two features.



on data is done in such a way that each PC is ordered in decreasing order by data variance.

Dimensionality reduction can be achieved by selecting the first n PCs which adequately represent the data set. The choice of n is often an *ad hoc* process (YEUNG; RUZZO, 2000). For visualization the two or three first PCs are usually chosen (EVERITT et al., 2011). The first two PCs can give suitable two-dimensional representation to view the cluster structure if a clear structure exists (JOLLIFFE, 2002). Figure 4.19 shows a scatterplot matrix with the estimate of data density (without volunteers who collaborate just for one day) from the three first PCs. It does not point out a clear evidence of clusters, but just a region with high density of points.

Figure 4.19 - Scatter plot matrix for the three fisrt PCs with density estimation (excluding data for volunteers who joined the project for just one day).



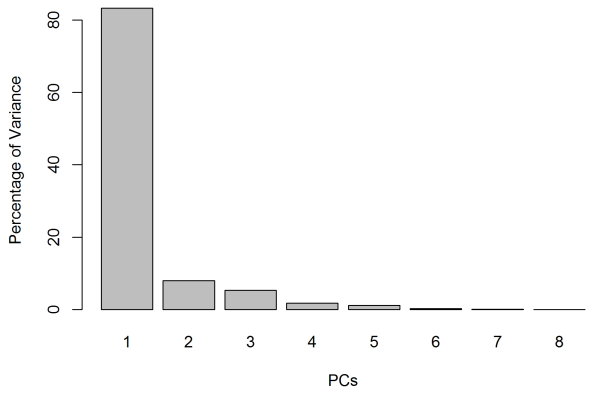
The quality of a projection can be quantified through the concept of dispersion (JOLLIFFE, 2002). A cloud of points of n dimensions is better represented into two dimensions if there is a high variability (variance) on these two dimensions. The percentage of variance is also the percentage of information explained by a given dimension (HUSSON et al., 2010). Figure 4.20 shows the distribution of variance on each PC. Note that the first two PCs represent more than 90% of information about the data distribution.

Factors like differences on magnitude and the existence of variables with irrelevant information may hide cluster structure. Next, we investigate two approaches which may help improve the detection of groups on data: standardization in section 4.3.1 and the relevance of features on grouping data in section 4.3.2.

4.3.1 Data Standardization

In many applications of cluster analysis the data are not used directly, they are transformed by some process, such as standardization or normalization. There is not a consensus about the use of the terms standardization and normalization. This work uses the term standardization as described by Milligan e Cooper (1988). Stan-

Figure 4.20 - Distribution of Variance on each PC.



dardization may be defined as a process whose purpose is to equalize the size (or magnitude) and the variability of the input data, giving all variables the same importance. Differences in the magnitude or scales of input data either hide cluster structure or in some cases may represent information that defines the clusters, so that, standardization may or not may be useful in a particular application (MILLIGAN; COOPER, 1988).

For convenience, let the matrix $n \times d$ denote a d-dimensional dataset. This matrix is given by:

$$(x_1^*, x_2^*, \ldots, x_n^*)^T = \begin{bmatrix} x_{11}^* & x_{12}^* & \ldots & x_{1d}^* \\ x_{21}^* & x_{22}^* & \ldots & x_{2d}^* \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}^* & x_{n2}^* & \ldots & x_{nd}^* \end{bmatrix}$$

Gan et al. (2007) summarize the various standardization methods as being the choice of different L_j (a location measure) and M_j (a scale measure) in equation 4.1. Table 4.1 shows some options of L_j and M_j.

$$x_{ij} = \frac{x_{ij}^* - L_j}{M_j} \tag{4.1}$$

The existence of numerous approaches hamper the decision process of which one to

[1]Uncorrected Standard Deviation

Table 4.1 - Some data standardization methods and its respective L_j and M_j according to Gan et al. (2007).

Name	L_j	M_j
z-score	$\bar{x}_j^*$	σ_j^*
USTD[1]	0	σ_j^*
Maximum	0	$_{1\leq i\leq n}\max(x_{ij}^*)$
Mean	$\bar{x}_j^*$	1
Range	$_{1\leq i\leq n}\min(x_{ij}^*)$	$_{1\leq i\leq n}\max(x_{ij}^*) -_{1\leq i\leq n} \min(x_{ij}^*)$
Sum	0	$\sum_{i=1}^{n} x_{ij}^*$

use (MILLIGAN; COOPER, 1988). In order to assess the possible improvements standardization methods may bring, we process the data through the following equations:

$$x_{ij} = \frac{x_{ij}^* - \bar{x}_j^*}{\sigma_j^*} \tag{4.2}$$

$$x_{ij} = \frac{x_{ij}^*}{\sigma_j^*} \tag{4.3}$$

$$x_{ij} = \frac{x_{ij}^*}{_{1\leq i\leq n}\max(x_{ij}^*)} \tag{4.4}$$

$$x_{ij} = \frac{x_{ij}^*}{_{1\leq i\leq n}\max(x_{ij}^*) -_{1\leq i\leq n} \min(x_{ij}^*)} \tag{4.5}$$

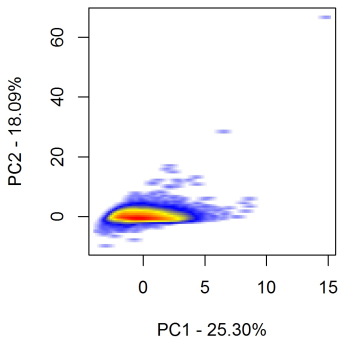
$$x_{ij} = \frac{x_{ij}^* -_{1\leq i\leq n} \min(x_{ij}^*)}{_{1\leq i\leq n}\max(x_{ij}^*) -_{1\leq i\leq n} \min(x_{ij}^*)} \tag{4.6}$$

$$x_{ij} = \frac{x_{ij}^*}{\sum_{i=1}^{n} x_{ij}^*} \tag{4.7}$$

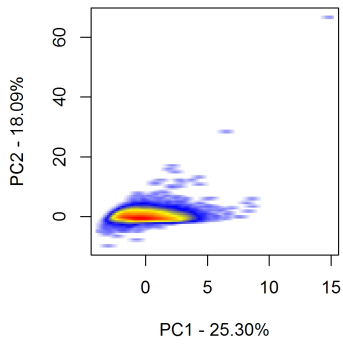
Using these equations, the original dataset (S_0) was transformed resulting in six new datasets (S_1 to S_6). Figure 4.21 shows the scatterplot of projection of the first two principal components for each dataset (S_1 to S_6) obtained through the six methods of

standardization. This experiment suggests at least two regions with higher density of points (see the projection of S_3, S_4 or S_5 in Figure 4.21). The projection of the dataset S_1 (Figure 4.21(a)) and S_2 (Figure 4.21(b)) and S_6 (Figure 4.21(f)) did not evidence clusters. On the other hand, the projection of the dataset S_3 (Figure 4.21(c)), S_4 (Figure 4.21(d)) and S_5 (Figure 4.21(e)) showed two regions (in red) with high density of points.

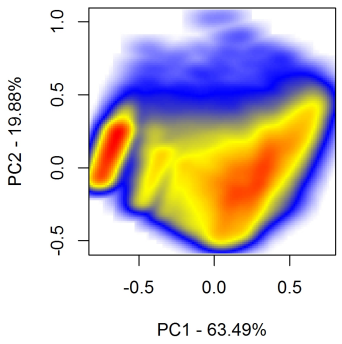
Figure 4.21 - Scatterplot of the first two principal components of dataset standardized by the six methods of standardization.



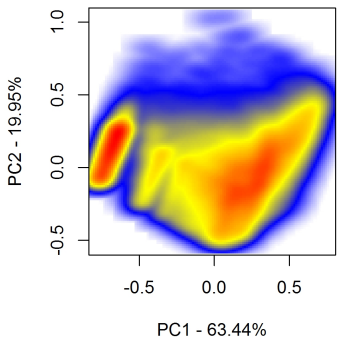
(a) S_1 - 43.39% of information



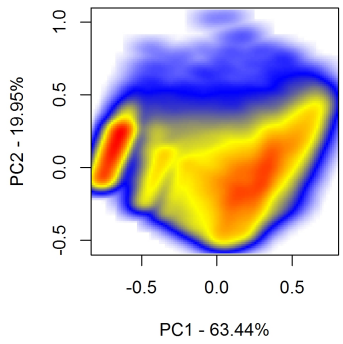
(b) S_2 - 43.39% of information



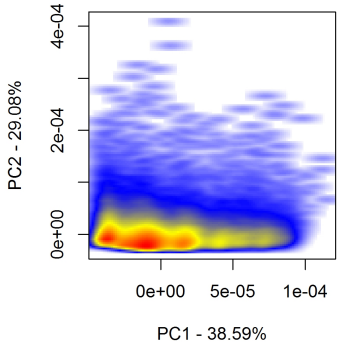
(c) S_3 - 83.38% of information



(d) S_4 - 83.39% of information



(e) S_5 - 83.39% of information



(f) S_6 - 67.67% of information

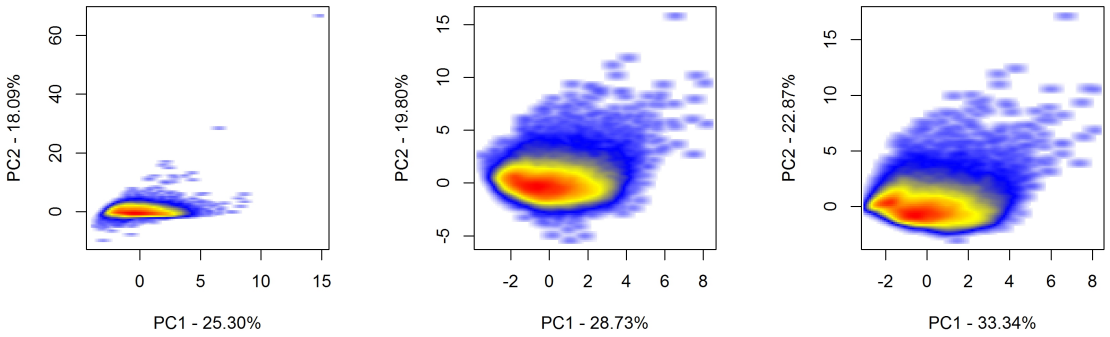
4.3.2 Relevance of Features on Grouping Data

In a previous work (MORAIS; SANTOS, 2015) we showed that the features Relative Active Duration, Assiduity and Distribution of Collaboration were able to describe some behavioral aspect of volunteers. In this study we proposed the addiction of five new features (i.e. Maximum Sustained Effort, Processing Power, Novelty, Challenge and Recurrence) based on model for volunteer interaction with web-based citizen science projects (described in section 3.3.1).

In the following experiment we evaluate if the removal of these additional features brought any new evidence of groups within data. Starting with the entire dataset, we applied the PCA and removed the feature with the lowest coefficient related to the first PC. In order to evaluate visually the result of this removal we plot the first two PCs. The experiment stopped when any of the coefficient related to the features Relative Active Duration, Assiduity or Distribution of Collaboration was the lowest.

To perform this experiment we selected three datasets: S_1, S_3 and S_6, where their projection are shown in Figure 4.21(a), Figure 4.21(c) and Figure 4.21(f), respectively. We make this choice because in Figure 4.21 these datasets yield visually different results on data distributions. It is noteworthy that the projection of the datasets S_2 (Figure 4.21(b)), S_4 (Figure 4.21(d)) and S_5 (Figure 4.21(e)) are similar to S_1 or S_3.

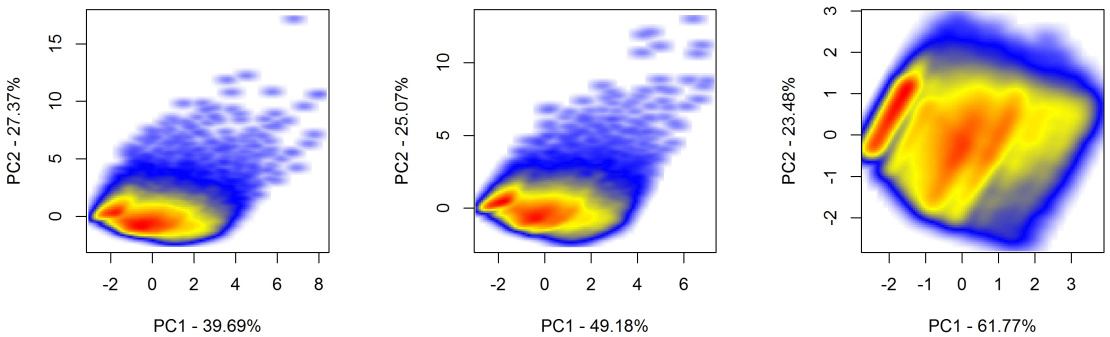
Figure 4.22 - Scatterplot of the first two principal components of dataset S_1 on each step of experiment.



(a) Complete dataset (8 features)

(b) After removing Novelty (7 features)

(c) After removing Challenge (6 features)



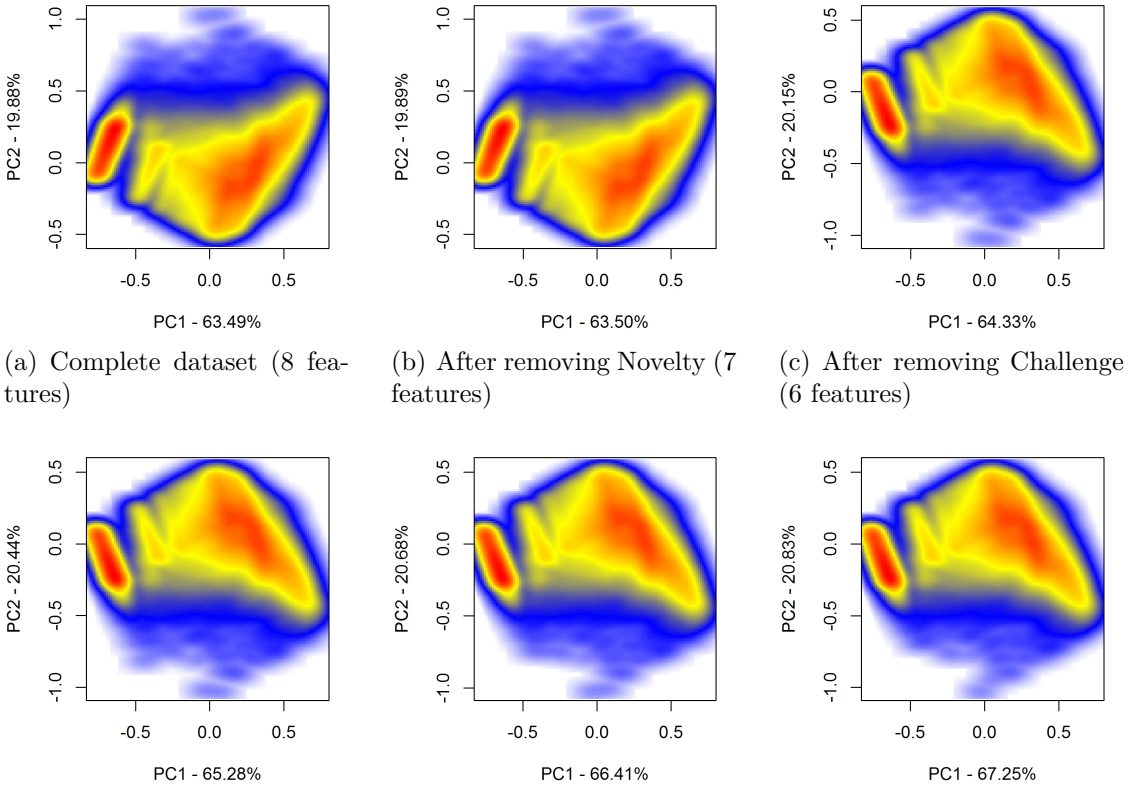
(d) After removing Processing Power (5 features)

(e) After removing Recurrence (4 features)

(f) After removing Maximum Sustained Effort (3 features)

Figure 4.22 presents the projection of dataset S_1 on each step of experiment. In the first step, Figure 4.22(a) shows a scatterplot of the first two PC of dataset S_1 containing all features. Note that there is no evidence of clusters. The removal of feature Novelty (Figure 4.22(b)) spread the data. However, it still did not show clusters. The initial evidence of clusters starts in Figure 4.22(c) with the removal of feature Challenge in the third step. The removal of feature Processing Power (Figure 4.22(d)) did not affect the visual appearance of previous step. In the fifth step, the removal of feature Recurrence shows the evidence of two groups separated by a tiny space in Figure 4.22(e). A clear evidence of two groups appeared in Figure 4.22(f) with the removal of feature Maximum Sustained Effort.

Figure 4.23 - Scatterplot of the first two principal components of dataset S_3 on each step of experiment.



(a) Complete dataset (8 features)

(b) After removing Novelty (7 features)

(c) After removing Challenge (6 features)

(d) After removing Processing Power (5 features)

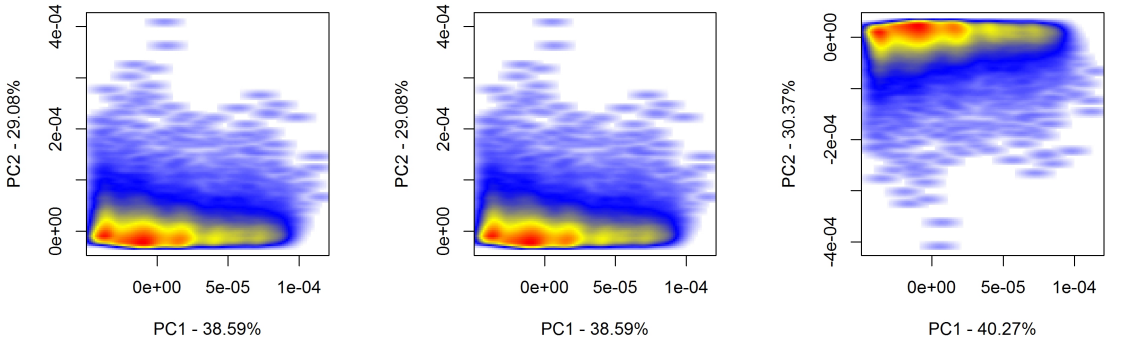
(e) After removing Recurrence (4 features)

(f) After removing Maximum Sustained Effort (3 features)

Figure 4.23 presents the experiments done with dataset S_3. Following the same process, the removal of features Novelty, Challenge, Processing Power, Recurrence and Maximum Sustained Effort did not affect the initial evidence of two groups on

the data. Note that the removal of feature Novelty in Figure 4.23(b), Challenge in Figure 4.23(c), Processing Power in Figure 4.23(d), Recurrence in Figure 4.23(e) and Maximum Sustained Effort in Figure 4.23(f) show the same data distribution.

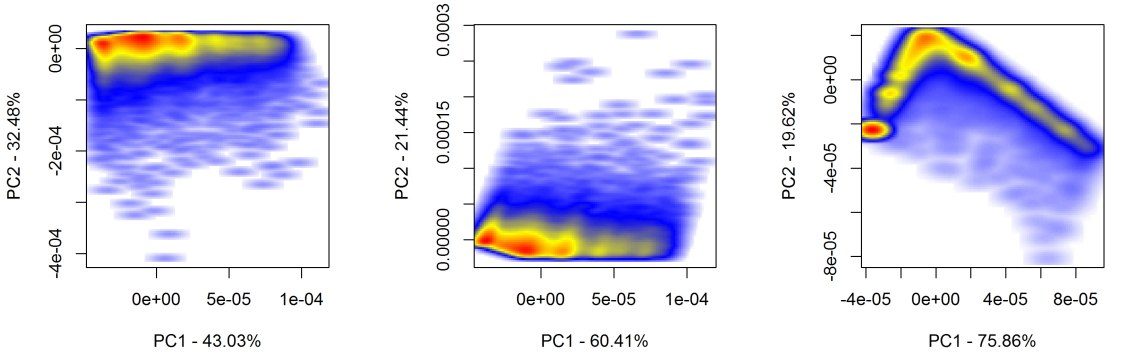
Figure 4.24 - Scatterplot of the first two principal components of dataset S_6 on each step of experiment.



(a) Complete dataset (8 features)

(b) After removing Novelty (7 features)

(c) After removing Challenge (6 features)



(d) After removing Recurrence (5 features)

(e) After removing Processing Power (4 features)

(f) After removing Maximum Sustained Effort (3 features)

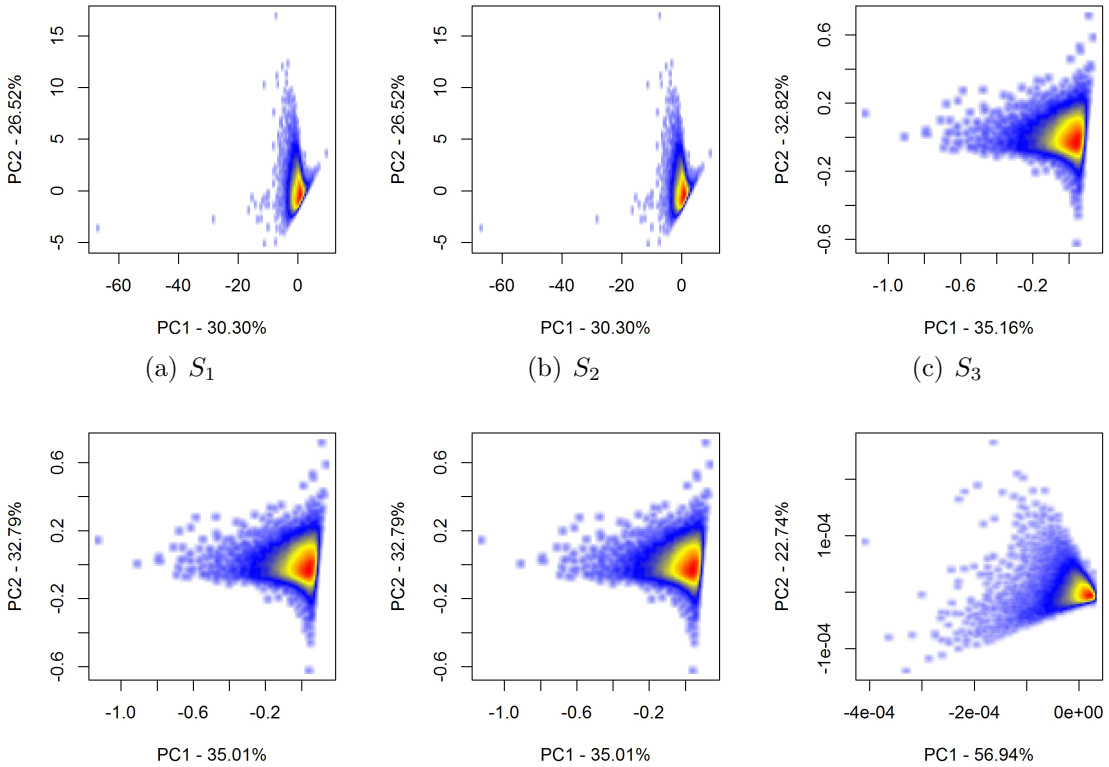
Figure 4.24 presents the steps of experiments done with dataset S_6. The removal of features Novelty, Challenge, Recurrence and Processing Power did not significantly affect the initial appearance of the data projection, see Figure 4.24(b), 4.24(c), 4.24(d) and 4.24(e), respectively. At the last step, the removal of feature Maximum Sustained Effort shows the evidence of two groups on data, see Figure 4.24(f).

These experiments allowed us to conclude that the features Relative Active Duration, Assiduity and Distribution of Collaboration were able to show groups which were hidden in the datasets. However, it should be noted that the standardization methods which resulting in datasets S_1 and S_6 could not give the same importance

(magnitude) for all variables. This fact may be the reason why these datasets did not show any clusters. The evidence that the five features may be not relevant for clustering analysis is showed by the experiments performed with the dataset S_3, see Figure 4.23. Initially, the dataset S_3 showed two regions with high density of points (in red). The removal of the five new features did not affect the data distribution, i.e., it is possible to observe these two regions using just the features Relative Active Duration, Assiduity and Distribution of Collaboration.

In order to verify the data distribution without the features Relative Active Duration, Assiduity and Distribution of Collaboration, we removed these features from the analyzed datasets (S_0 to S_6) and observed their projections. Figure 4.25 shows that there is no evidence of groups within these datasets without the features Relative Active Duration, Assiduity and Distribution of Collaboration.

Figure 4.25 - Scatterplot of the first two principal components of the datasets S_0 to S_6 without the features Relative Active Duration, Assiduity and Distribution of Collaboration.



(a) S_1 (b) S_2 (c) S_3

(d) S_4 (e) S_5 (f) S_6

5 CLUSTERING ANALYSIS

Cluster analysis is probably the preferred generic term for procedures which seek to uncover groups in data (EVERITT et al., 2011). In short, it may be described as a process organizing data into groups based only on information found in the data. Formally, these groups are called clusters and are defined in terms of internal cohesion (homogeneity) and external isolation (separation), so that a cluster is a collection of similar data objects within the same group and dissimilar to the objects in other groups.

In practice, the process of seeking groups in data involves deciding which data will be used to cluster, the selection of computational method (i.e. a clustering technique), the choice of a proximity measure to discern if two data objects are similar or dissimilar and how to evaluate the quality of the discovered groups. Each step of this process has a significant number of options to be chosen.

There are many applications of cluster analysis to practical problems. In the context of this dissertation, techniques of data clustering are applied in a dataset which describes the features of the volunteers' behavior in a citizen science project, in order to detect groups of volunteers which follow similar behavioral pattern. Throughout this chapter, we present the basic concepts of the cluster analysis literature in Section 5.1 and the experiments performed as well as the clustering techniques explored for this work in Section 5.2.

5.1 Basic Concepts

In most applications of cluster analysis, the data are organized in a multivariate matrix M_{nd}. This matrix represents n data objects with d variables. The rows and columns of the matrix represent different entities, each row corresponds to a data object while each column consists of a characteristic of the object, so that, the entry x_{ij} in M_{nd} gives the value of the jth variable (column) of the data object i (row). The set of possible values assumed by a variable is determined by its type. The data type has a significant impact on the process of cluster analysis (AGGARWAL; REDDY, 2013). It may determine the choice of a proximity measure and influence the selection of a clustering algorithm.

The types of variables that often occur in cluster analysis are: **Nominal** also called categorical, their values are symbols or names of things without any meaningful order, **Binary** that are nominal variables with only two categories (or states) 0

or 1, **Ordinal** which is similar to nominal variables, except that their values have a meaningful order and, **Numeric** that are measurable quantities represented in integer or real values which can be interval-scaled or ratio-scaled (HAN; KAMBER, 2006). The d variables of a multivariate matrix may be composed by data types of one or different types. In the context of this research, the data matrix is composed by numeric values which measure the aspects of how volunteers interact with web-based citizen science projects.

The central step to identify groups hidden in data is to know about how close individuals (data objects) are to each other, or how far apart they are (EVERITT et al., 2011). There is a wide number of measurements of proximity proposed in literature, each of these measurements were proposed to deal with one or more types of variables. Proximity is a generic term which refers to measures of dissimilarity, distance or similarity.

In general, clustering algorithms are based on measures of dissimilarity. Given a space of attributes, i.e. the set of points into a d dimensional space where d is the number of variables which describes a data object, a measure of dissimilarity is a function $f(i, j)$ which produces a real number indicating how far apart the data objects i and j are from each other. Details about the different kind of proximity measures are shown in (EVERITT et al., 2011). The most familiar dissimilarity measure for numerical variables used in cluster analysis is the *Euclidean* distance:

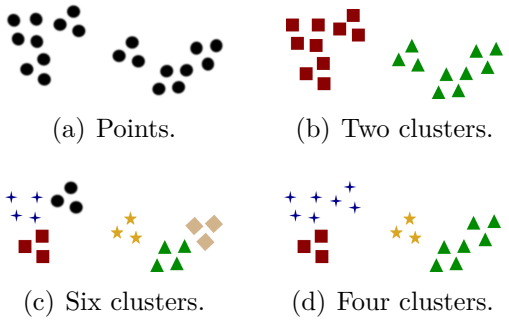
$$f(i,j) = \sqrt{\sum_{k=1}^{d}(i_k - j_k)^2}$$

In practice, for many applications the notion of a cluster is not well defined (TAN et al., 2006). The definition of a cluster is often imprecise and depends on the nature of data and the desired results. Moreover, distinct clustering methods, or even the same clustering method run with different input parameters, applied in a given dataset usually provide different groups of data. Figure 5.1 shows how the same data could lead to different clusters.

Consider the set of points in a two-dimensional space showed by Figure 5.1(a). Visually, these points seem to be divided into two groups as shown in Figure 5.1(b). Moreover, by an artifact of the human visual system, each of the two groups may be reorganized into three subclusters, see Figure 5.1(c). It may not be unreasonable to say that the points form four clusters as shown in Figure 5.1(d), however, for some

applications it could be the desired result (TAN et al., 2006). In short, data clustering is done in order to find useful groups of objects, where useful should be defined in terms of the goals of the data analysis.

Figure 5.1 - Different ways of clustering the same set of points.



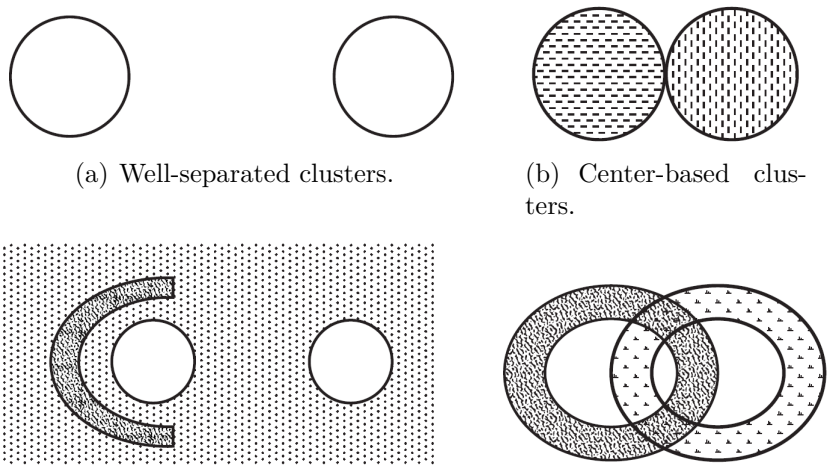
(a) Points. (b) Two clusters.

(c) Six clusters. (d) Four clusters.

SOURCE: Tan et al. (2006).

Practical problems have shown different types (or notions) of clusters. In order to illustrate this concept, Figure 5.2 shows distinct distributions of points which may represents different types of clusters.

Figure 5.2 - Types of clusters illustrated into two-dimensional points space.



(a) Well-separated clusters. (b) Center-based clusters.

(c) Density-based clusters. (d) Conceptual clusters.

SOURCE: Tan et al. (2006).

Figure 5.2(a) shows an example of well-separated set of objects in which each object is closer to every other object in the cluster than to any object not in the cluster. This type of cluster is an idealistic definition of cluster which is satisfied only if the data contains natural clusters (TAN et al., 2006). The shape of this kind of cluster does not need to be globular. In this kind of cluster given any two points in different groups, the distance between them is larger than the distance between any point within a group.

Figure 5.2(b) shows an example of center-based clusters, also called prototype-based. In a center-based clusters, each point is closer to the center (prototype) of its cluster than to center of any other cluster. For numerical data, the prototype is usually the average of all points in the cluster. In contrast to well-separated clusters, center-based clusters tends to be globular. Another example is shown in Figure 5.2(c), this type is called Density-Based. A cluster in this case is a dense region of objects that is surrounded by a region of low density. Clusters may also be defined by a set of objects that share some property. In conceptual clusters, points which are very close may conceptually belong to different clusters. Figure 5.2(d) exemplifies this concept.

5.1.1 A Categorization of Clustering Methods

There are many clustering methods and it is difficult to organize them into categories, once some methods may have features from several categories (HAN; KAMBER, 2006). A relatively organized picture of clustering methods is presented by Han e Kamber (2006) and Aggarwal e Reddy (2013) with some differences in terminology. The major fundamental clustering methods can be categorized into few clustering methodologies such as:

Distance-based Methods: are the simplest and most fundamental methodology of cluster analysis. They are used in a wide variety of scenarios, their implementation is easily compared to other clustering methods and they can be used with almost all data types, as long as an appropriate distance function is chosen (AGGARWAL; REDDY, 2013). These methods can be generally divided into two types: *Partitioning Methods* and *Hierarchical Methods*.

In *Partitioning Methods*, a given data set with n data objects is organized into the k partitions (or clusters). These methods require some background knowledge to specify the number k of partitions, which may be a drawback in some cases. Methods of this category adopt a hard clustering procedure or fuzzy clustering procedure. In the first, a data object is either assigned to or not assigned to a cluster while in the

latter, a data object is assigned to a membership function for each of the clusters. Fuzzy clustering procedures are more flexible and robust to deal with noisy and uncertain data than hard clustering procedures (CHI et al., 1996). It is noteworthy that partitioning methods work well to find clusters with normal distribution, but are inefficient to find clusters with complex shapes (HAN; KAMBER, 2006).

Hierarchical Methods organize the data objects into a tree at varying levels of granularity. These methods can be further classified as agglomerative or divisive, depending on whether the hierarchical group is formed through a bottom-up (merge) or a top-down (split) strategy respectively. Advantages of these methods are the representation of the data into a hierarchical form which is useful for data summarization and visualization to small or medium size datasets, moreover, they do not require background knowledge.

Density-based Methods: were proposed to deal with requirements of scalability to large datasets and the ability to detect and remove noise and outliers (AGGARWAL; REDDY, 2013). Methods of these approaches assume that the density within areas of noise is lower than the density in any of the clusters. Therefore, areas of low density are treated as noise and are not assigned to any cluster. These methods can be considered as a nonparametric approach, with no assumptions about the shape, the number of clusters or the data distribution. This kind of method is a good option when data contains clusters which are irregular or intertwined, and when noise and outliers are present (TAN et al., 2006). Methods of this category have some key design questions to be answered as how the density is estimated and how the connectivity is defined. Some examples of methods are DBSCAN (ESTER et al., 1996), DENCLUE (HINNEBURG; KEIM, 1998) and OPTICS (ANKERST et al., 1999).

Grid-based Methods: are closely related to density-based approach (AGGARWAL; REDDY, 2013). Some authors describe them as being a specific class of density-based methods. In a general approach, these methods partition the data into a finite number of cells arranged in a grid structure and form clusters through the cells. In these methods, clusters correspond to regions that are more dense in data points than their surroundings. The main advantage of these approaches is their fast processing time (HAN; KAMBER, 2006). The computational complexity of their algorithms is typically independent of the number of data objects, but dependent on the number of cells. The efficiency of grid-based clustering algorithms comes from how data points are grouped into cells and clustered collectively rather than individually (AGGARWAL; REDDY, 2013). Some examples of methods are STING (WANG et al., 1997),

WaveCluster (SHEIKHOLESLAMI et al., 1998) and CLIQUE (AGRAWAL et al., 1998).

Model-based Methods: attempt to optimize the fit between a given data and some mathematical model (HAN; KAMBER, 2006). These methods are composed of different approaches like Probabilistic and Generative Methods and Neural Networks. In probabilistic models, for example, the methods are often based on the assumption that the data are generated by a mixture of underlying probability distributions. The clustering problem is transformed into a parameter estimation problem (AGGARWAL; REDDY, 2013). The neural network approach is motivated by biological neural networks. It tends to represent each cluster as an example (or prototype) achieved by a numerical process called learning.

5.1.2 Evaluation of Clustering

Blindly applying a clustering method on a given data set will return clusters, but these clusters may not be meaningful or may be misleading. Therefore, it is important to assess the quality of the results generated by a clustering method (HAN; KAMBER, 2006). Measuring clustering quality has long been recognized as one of the essential issues to the success of clustering applications (JAIN; DUBES, 1988). However, there is no conclusive solution to cluster validation, even though much effort has been spent on this problem (AGGARWAL; REDDY, 2013).

In the literature, a wide variety of clustering validation measures have been proposed. These measures can be categorized into external clustering validation and internal clustering validation (AGGARWAL; REDDY, 2013). On external validation measures, information provided by human experts is used for clustering validation. In practice, some real applications do not have external information available and only internal measures are possible. On internal clustering validation only the data and the results of cluster algorithms are used for clustering validation. It is noteworthy that different validation measures may be appropriate for different clustering algorithms (AGGARWAL; REDDY, 2013). Experiments done by Xiong (AGGARWAL; REDDY, 2013) show that internal clustering validation may be affected by data aspects like noise, density, sub-clustering, skewed distribution and arbitrary shape.

5.2 Cluster Analysis of Volunteer Data

Recent studies have analyzed usage logs (records of who did what and when) in an attempt to infer information about volunteers' motivation. These studies focus on detection of groups of volunteers whose interactions with the project follow a

similar behavior pattern. In this context, cluster analysis is a field of knowledge which provides a wide number of techniques to support the detection (or uncover) of groups of volunteers in data.

It is worth mentioning that some volunteers' behaviors are well know and easily detectable. Often volunteers on web-based citizen science projects contribute to these projects in small quantities or in short bursts (EVELEIGH et al., 2014). Previous analyses of usage logs have showed the existence of two main groups of volunteers. One is formed by volunteers who contribute just for one day, and the other is composed by volunteers who contribute at least for two days. The existence of these two main groups were noted by the analyses of the usage logs of Galaxy Zoo (MAO et al., 2013; MORAIS et al., 2013; PONCIANO et al., 2014), The Milky Way (PONCIANO et al., 2014), the Sun4All (PONCIANO, 2015) and Cell Spotting (PONCIANO, 2015).

Volunteers who collaborate just for one day have been considered as a volunteer profile and were removed from cluster analysis by previous works like Ponciano e Brasileiro (2015) and Ponciano (2015). Following this line, we removed these volunteers from the data to be clustered. Therefore, the experiments described in this section were performed with 51,512 volunteers who performed 69,532,456 collaborations.

This section aims to present the challenges and issues of applying clustering techniques to detect groups of volunteers with similar pattern behavior. The experiments described in this section include K-Means, Fuzzy C-Means, DBSCAN and Self-Organizing Maps. In order to verify the possible benefits of standardization of variables, the chosen clustering methods were applied on the original data set (S_0) and on six datasets (S1 to S6) obtained through the methods of standardization described on section 4.3.1. Moreover, the experiments are organized into two parts. One considering just the features Relative Active Duration, Assiduity and Distribution of Collaboration and the other considering the entire set of features, i.e., these three features and the other five: Maximum Sustained Effort, Processing Power, Novelty, Challenge and Recurrence. The following sections present in details the experiments performed with those clustering techniques.

5.2.1 Partitioning Methods: K-Means and Fuzzy C-Means

One of the most used method of partitioning is K-Means algorithm and its variations. The basic K-Means steps are described by the Algorithm 5.2.1. In short, given an input dataset and the number k of partitions, K-Means randomly selects k

data objects. Each one of the k data objects initially represents a cluster mean, also called centroid. For each data object within the dataset, a data object is assigned to a cluster that is the most similar based on the distance (often Euclidean distance) between this data object and the cluster mean (the centroid). The algorithm then updates the centroid of each cluster by calculating its new cluster mean. This process iterates until the centroids do not change significantly, which happens when a criterion function converges. Typically, the criterion consists of minimizing the Square-Error function:

$$E = \sum_{i}^{k} \sum_{p \epsilon C_i} \|p - m_i\|^2 ,$$

where E is the sum of the square error for all data objects into the dataset; p is the point representing a given data object; and m_i is the mean of cluster C_i. This criterion tries to find k clusters which are as compact and as separated as possible.

Algorithm 5.2.1: Basic K-Means algorithm

Input: The number k of clusters

Result: A set of k clusters

1 Select k points as initial centroids;
2 **repeat**
3 | From k clusters by assigning each point to its closest centroid;
4 | Recompute the centroid of each cluster;
5 **until** *Centroids do not change*;

Previous works on characterization of volunteers (PONCIANO; BRASILEIRO, 2015; PONCIANO, 2015) already used K-Means to extract groups of volunteers. Although K-Means has been helpful to uncover groups of volunteers in these works, this algorithm is inefficient in some scenarios. In the clustering literature, K-Means is known as being good for finding cluster with normal distribution and it is inefficient to deal with data that doesn't contain compact and well-defined cluster. Moreover, the results obtained by K-Means are strongly dependent on the initial centroids and the algorithm is sensitive to noise and outliers, i.e. a small number of noise or outliers may substantially influence the update of centroids (HAN; KAMBER, 2006).

In a hard clustering procedure, like K-Means, a data object is either assigned to or not assigned to a cluster. In practice, there are cases where a data object may be assigned to one cluster as well as to another. In scenarios were the clusters were not completely disjointed, a fuzzy clustering procedure may provide better results than hard clustering procedures. In a fuzzy clustering procedures the data objects

are organized into a given number of partitions through membership function. This function indicates the degree of membership of a data object to a given cluster, i.e., each data object belongs to all clusters with varying degrees of membership between 0 and 1, where values close to one indicate higher confidence in the assignment of the pattern to the cluster.

The Fuzzy C-Means algorithm is the most popular fuzzy clustering procedure (CHI et al., 1996). This algorithm is considered an extension of the K-Means algorithm. The basic Fuzzy C-Means steps are described by the Algorithm 5.2.2. The algorithm may be considered as an iterative optimization procedure whose objective function is the sum of the squared Euclidean distance between each data object and its corresponding centroids, with the distance weighted by the fuzzy memberships. In each step, the algorithm aims to minimize the function:

$$J(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \left\|x_k - v_i\right\|^2,$$

where x_k is the point representing a given data object and v_i a cluster center; $V = v_1, ..., v_c$ a set of c clusters centers; m an exponent weight factor; and U a $c \times n$ matrix, where u_{ik} is the ith membership value of the kth data object x_k.

Algorithm 5.2.2: Basic Fuzzy C-Means algorithm

Input: The number of clusters (c) and a value of $m \epsilon [1, \infty)$

Result: A set of c clusters

1 Initialize the membership values $U^{(0)}$ randomly or based on an approximation;
2 Select c points as initial centroids, i.e., initialize the cluster centers $V^{(0)}$;
3 **repeat**
4 Given the membership values $U^{(\alpha)}$, compute the cluster centers $V^{(\alpha)}$ according to Equation 5.1;
5 Update the membership values $U^{(\alpha)}$ according to Equation 5.2;
6 **until** $max|u_{ik}^{\alpha} - u_{ik}^{\alpha-1}| < \varepsilon$;

In short, given an input dataset with n data objects, c the number of partitions and m a weight factor, Fuzzy C-Means initializes the membership values of matrix U and randomly selects c data objects to represent the initial clusters. For each data object within the dataset, V (the cluster centers) and U (the membership values) are updated by the equations 5.1 and 5.2, respectively.

$$v_i = \frac{1}{\sum_{k=1}^{n} u_{ik}^{m}} \sum_{k=1}^{n} u_{ik}^{m} x_{ik} \tag{5.1}$$

$$u_{ik} = \frac{\left[\frac{1}{|x_k - v_i|^2}\right]^{\frac{1}{(m-1)}}}{\sum_{j=1}^{c}\left[\frac{1}{|x_k - v_j|^2}\right]^{\frac{1}{(m-1)}}} \tag{5.2}$$

The common characteristic of any partitioning methods is to require a number of partitions (as an input parameter) in which the data objects will be organized. Applying these methods without the most appropriate number of partitions returns clusters which may not be meaningful or may be misleading. The number of partitions is often estimated through techniques of cluster evaluation, i.e. validation measures, or specified by some background knowledge. In the first approach, this number is estimated by applying the method several times, each time with a different number of partitions and choosing the best number based on some measure that evaluates the clustering results. Different validation measures may be appropriate for different clustering algorithms (AGGARWAL; REDDY, 2013).

Silhouette analysis is one of the several techniques described in the clustering literature which may measure the quality of the resulting clusters (ROUSSEEUW, 1987). It is defined as being a function $s(i)$ which measures how well a data object i matches with a given cluster. With a range of -1 to 1, it is interpreted as following: values close to -1 indicates that i should be assigned to some neighboring cluster, while values close to 1 means that it is already in the appropriate cluster and, values close to 0 suggest that i is on the border of two or more natural clusters. The average of the $s(i)$ (called silhouette coefficient) for all objects i in the data set is used as a measure for cluster evaluation. Silhouette coefficients may be interpreted subjectively as shown in Table 5.1.

Table 5.1 - Rule of thumb to interpret the silhouette coefficient (STRUYF et al., 1997).

SC	Interpretation
$0.71 - 1.00$	A strong structure has been found.
$0.51 - 0.70$	A reasonable structure has been found.
$0.26 - 0.5$	The structure is weak and could be artificial.
≤ 0.25	No substantial structure has been found.

Another example of validation measure for hard clustering procedures is the Error Sum of Squares, also called Within-Groups Sum of Squares. In short, the Error Sum of Squares measures the difference between the centroid and its points, so

that, low values show how well the centroids represent the points assigned to them. Considering the Error Sum of Squares and the silhouette coefficient, the choice of k partitions consists of looking for the higher value of silhouette coefficient and the lower value of the Error Sum of Squares.

For fuzzy clustering procedures like Fuzzy C-Means, there are several techniques for cluster evaluation. The commonly used are Partition Coefficient, Partition Entropy, Fukuyama Sugeno and Xie Beni (SAAD; ALIMI, 2012). The appropriate number is usually found by applying the algorithm several times with different number of partitions, and seeking for a result which maximize Partition Coefficient and minimizes the others (SAAD; ALIMI, 2012). Besides the number of partitions, Fuzzy C-Means algorithm also requires a weight factor m as an input parameter. This parameter reduces the influence of small membership values, in a way that the larger the value of m, the smaller the influence of data objects with small membership values (CHI et al., 1996).

Next, the results and analysis of the K-Means and Fuzzy C-Means are described in details. For all experiments the algorithms were run with number of partitions varying from 2 to 30.

Results and Analysis of experiments performed with K-Means:
The best values for the silhouette coefficient for the datasets (S_0 to S_6), containing only the three features, are shown in Table 5.2.

Table 5.2 - The best values for silhouette coefficient obtained from the datasets with the features Relative Active Duration, Assiduity and Distribution of Collaboration.

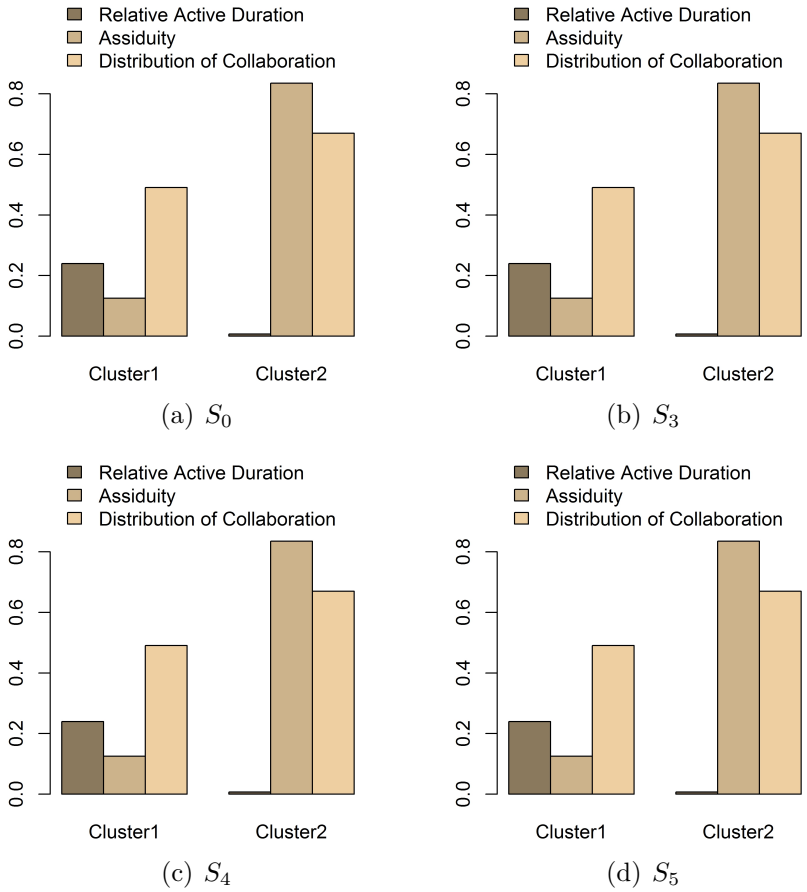
Data	Max(SC)	K
S_0	0.48	2
S_1	0.43	3
S_2	0.43	3
S_3	0.48	2
S_4	0.48	2
S_5	0.48	2
S_6	0.44	3

Silhouette coefficient indicated two clusters on the experiments done with datasets S_0, S_3, S_4 and S_5, and three clusters for datasets S_1, S_2 and S_6. However, according

to the rule of thumb (Table 5.1), these clusters could be artificial.

Figure 5.3 shows the values of the centroids obtained from experiments done with $k = 2$ and datasets S_0, S_3, S_4 and S_5. Observe that the clusters found from different datasets are similar.

Figure 5.3 - The values of centroids obtained from K-Means for $k = 2$ and the datasets S_0, S_3, S_4 and S_5.



(a) S_0 (b) S_3

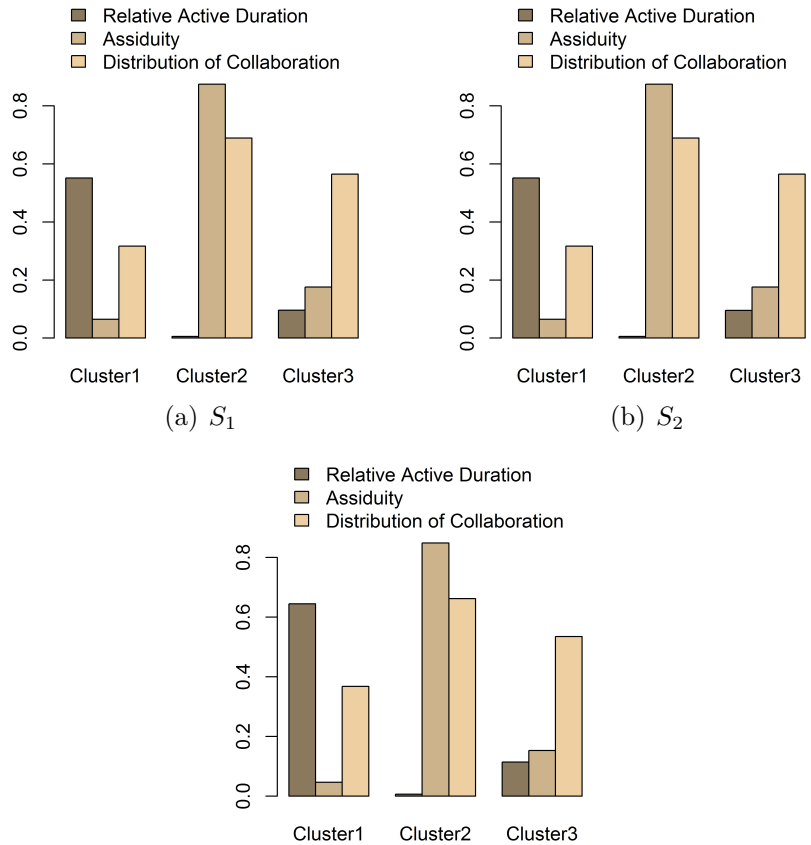
(c) S_4 (d) S_5

Consider Figure 5.3(a) to illustrate the two clusters. The analysis of the centroids suggests that in general volunteers performed most of the collaborations in one day (see that the value of the feature Distribution of Collaboration is higher than 0.5 in both clusters). The bar-plot labeled as *Cluster1* shows that the feature Relative Active Duration is close to 0.2, indicating that there are volunteers who joined the project for almost four months (i.e. 0.2×600 days $= 120$ days). Volunteers from this cluster were not assiduous (note that the feature Assiduity is close to 10%), so that

in practice they collaborated for some weeks (0.1×120 days $= 12$ days). The other cluster, (see the bar-plot labeled as *Cluster2*) indicates that there are volunteers who joined the project and abandoned it shortly afterwards, being assiduous during this period.

Figure 5.4 shows results of experiments performed with datasets S_1, S_2 and S_6 with plots of centroids for $k = 3$. Note that the clusters found from these datasets were similar.

Figure 5.4 - The values of centroids obtained from K-Means for $k = 3$ and the datasets S_1, S_2, S_6.



(a) S_1 (b) S_2

(c) S_6

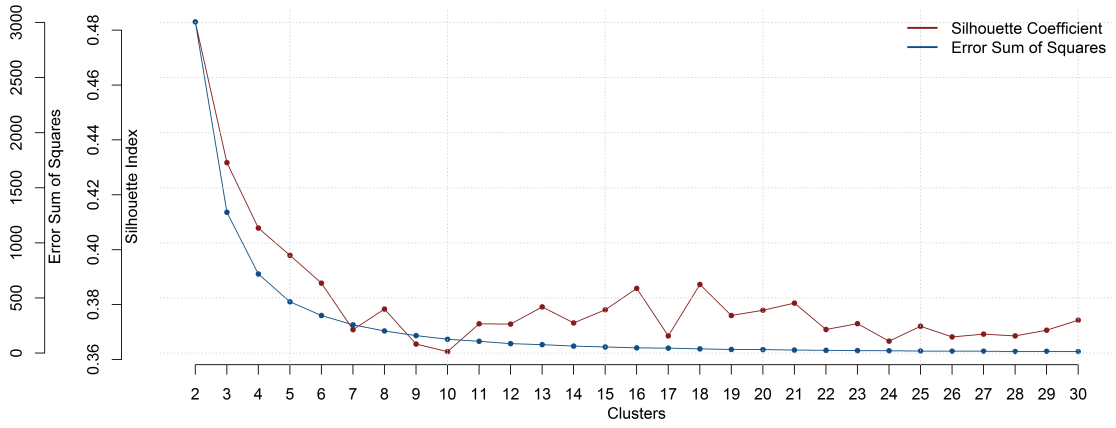
Consider Figure 5.4(a): one cluster is formed by volunteers who joined the project for two months (0.1×600 days $= 60$ days), but in practice collaborated just for some weeks (0.2×60 days $= 12$ days) and performed almost all collaborations in one day; the second cluster consists of assiduous volunteers who joined the project and abandoned it shortly afterwards, performing almost all collaborations in one

day; the third cluster is formed by volunteers who joined the project for almost one year (0.6×600 days $= 360$ days), but were not assiduous and collaborated just for about one month (0.1×360 days $= 36$ days). In contrast to other clusters, volunteers from the third cluster distributed their collaborations in a most homogeneous way along the active days.

These experiments did not indicate the presence of assiduous volunteers who joined the project for years and spread their collaboration along their active days. In Chapter 4, exploratory data analysis (section 4.2) singled out the evidence of this kind of behavior. Considering the volunteers who joined the project for more than one year (i.e. the feature Relative Active Duration is higher than or equal to 0.6) and defining that these volunteers can be called assiduous if the feature Assiduity is higher than 0.6, manual analysis of the data indicates at least eighteen volunteers who follow this behavior. As an attempt to detect this profile and investigate the existence of other patterns, some experiments considering the Error Sum of Squares were done with K-Means in order to evaluate the choice of other values to k.

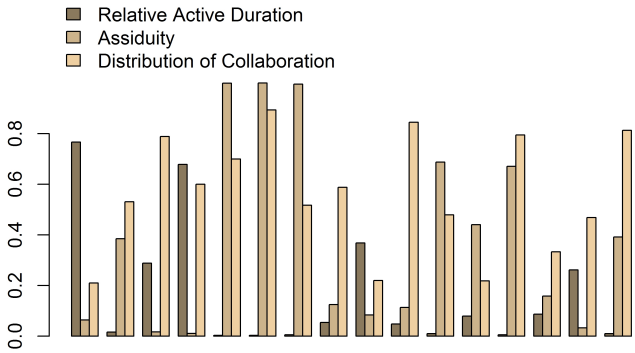
Figure 5.5 shows the values of the Error Sum of Squares and silhouette coefficient for each partition obtained by the experiments performed with dataset S_0. Looking for the higher value of silhouette coefficient and the lower value of the Error Sum of Squares, the better options for the value of k are $k = 16$ or $k = 18$.

Figure 5.5 - Values of the Error Sum of Squares and silhouette coefficient obtained by the experiment performed with dataset S_0.

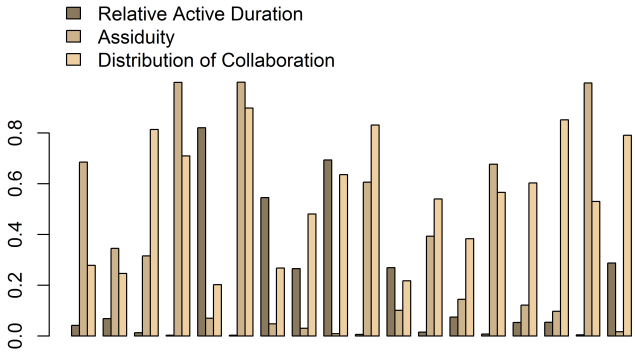


The clusters obtained through $k = 16$ and $k = 18$ are illustrated in Figure 5.6. Note that Figure 5.6 does not show clusters whose features Relative Active Duration and Assiduity are higher than 0.6 which would represent the volunteers singled out during the exploratory data analysis. Moreover, the clusters shown by Figures 5.6(a) and Figures 5.6(b) did not point out new evidence of volunteers' profiles. The clusters could be described as slight variations of the profiles already described. Experiments done with datasets S_1 to S_6 point out the same results.

Figure 5.6 - Groups of volunteers obtained from K-Means applied on datasets S_0 for $k = 16$ and $k = 18$.



(a) $k = 16$



(b) $k = 18$

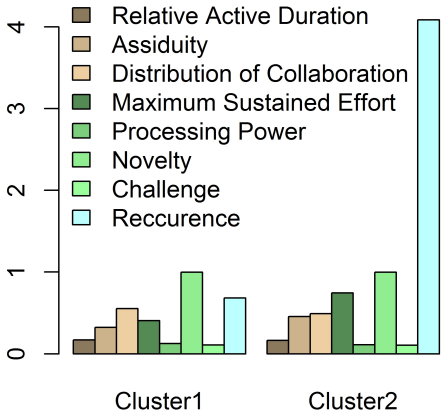
Table 5.3 shows the best values for the silhouette coefficient for each dataset (S_0 to S_6) composed by the eight proposed features. According to the rule of thumb, the results indicate that no substantial structure could be found with exception of dataset S_0. Silhouette coefficient suggests two clusters into the dataset S_0.

Table 5.3 - The best values of Silhouette Coefficient achieved from K-Means applied on all features

Data	Max(SC)	K
S_0	0.62	2
S_1	0.13	3
S_2	0.13	2
S_3	0.17	2
S_4	0.17	2
S_5	0.17	2
S_6	0.20	3

The main difference between these clusters is the feature Recurrence, the values of the other features are similar in both clusters, see Figure 5.7. One cluster is composed by volunteers who usually did not return to collaborate on the same active day and the other is composed by volunteers who returned on average of four times.

Figure 5.7 - The centroid obtained from K-Means for $k = 2$ considering the silhouette coefficient and the dataset S_0.



With regard to the value of the features Relative Active Duration, Assiduity and Distribution of Collaboration the clusters show volunteers who collaborated for almost four months without being assiduous and performed a good amount of collaboration just in one day. The other features indicate that these volunteers maintaining their mental context for about thirty minutes, spent some seconds thinking about the

tasks, did not perform a task more than one time in the same period of engagement and had to deal only with easy tasks.

Considering the Error Sum of Squares, Figure 5.8 shows the values of silhouette coefficient and the Error Sum of Squares, for k from 2 to 30. Note, in Figure 5.8, that with the decrease of the Error Sum of Squares, the silhouette coefficient achieves values lower than 0.25. According to the rule of thumb, silhouette coefficient values lower than 0.25 indicates that no substantial structures were found. The red dashed line in Figure 5.8 delimits the values of k whose silhouette coefficient is higher than 0.25.

Figure 5.8 - Values of the Error Sum of Squares and silhouette coefficient obtained by the experiment performed with dataset S_0 containing the eight features.

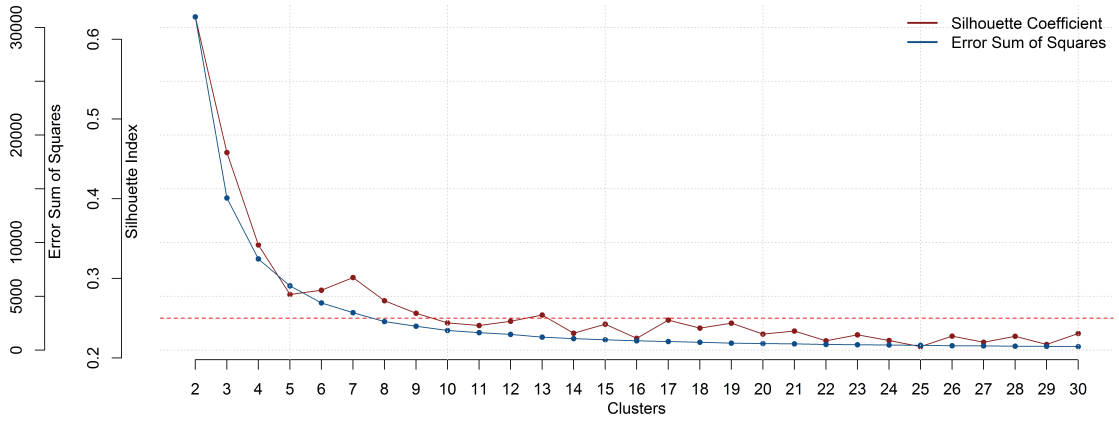
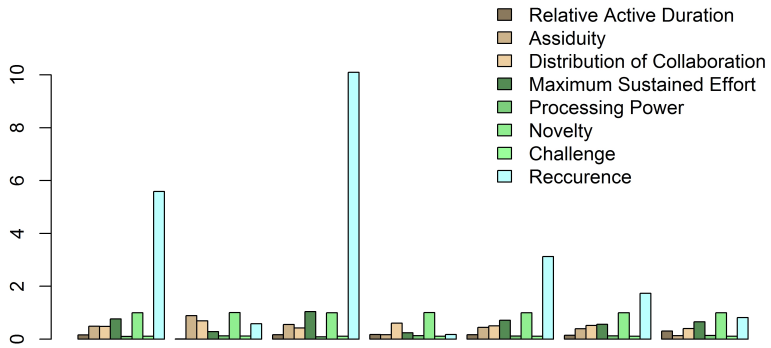


Figure 5.9 shows the centroids for $k = 7$, $k = 9$ and $k = 13$. With regard to the value of the features Relative Active Duration, Assiduity and Distribution of Collaboration, the analysis of the centroids did not point out new evidences of volunteers' profiles with respect to the first experiments performed with three features. The other five features present few variations between the clusters with exception of the feature Recurrence as observed during the analysis of the two clusters pointed out by silhouette coefficient showed in Figure 5.7.

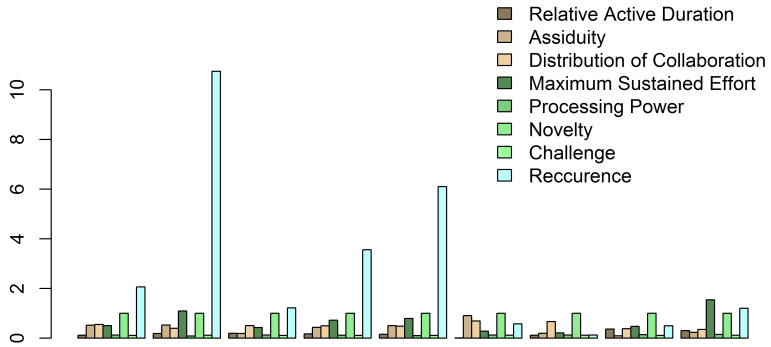
Results and Analysis of experiments performed with Fuzzy C-Means:
Figure 5.10 shows the cluster validation measures for c varying from 2 to 30, obtained through the dataset S_0 for $m = 2$ and considering the features Relative Active Du-

ration, Assiduity and Distribution of Collaboration. These measures do not indicate a clear evidence of clustering structure into data, i.e., it is not possible to choose a value of c which maximizes Partition Coefficient and minimizes Partition Entropy, Fukuyama Sugeno and Xie Beni.

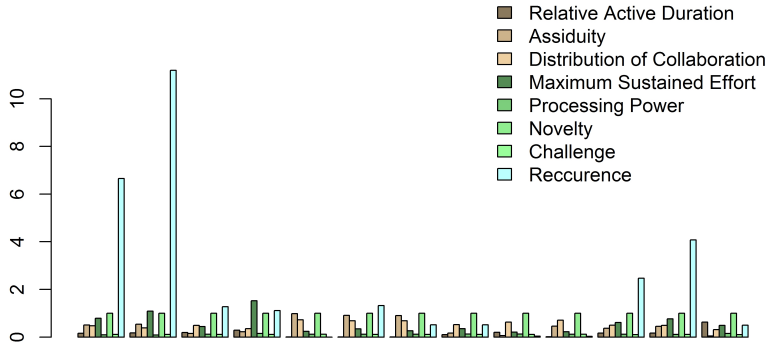
Figure 5.9 - The centroids from K-Means considering the silhouette coefficient, the Error Sum of Squares and the dataset S_0.



(a) $k = 7$

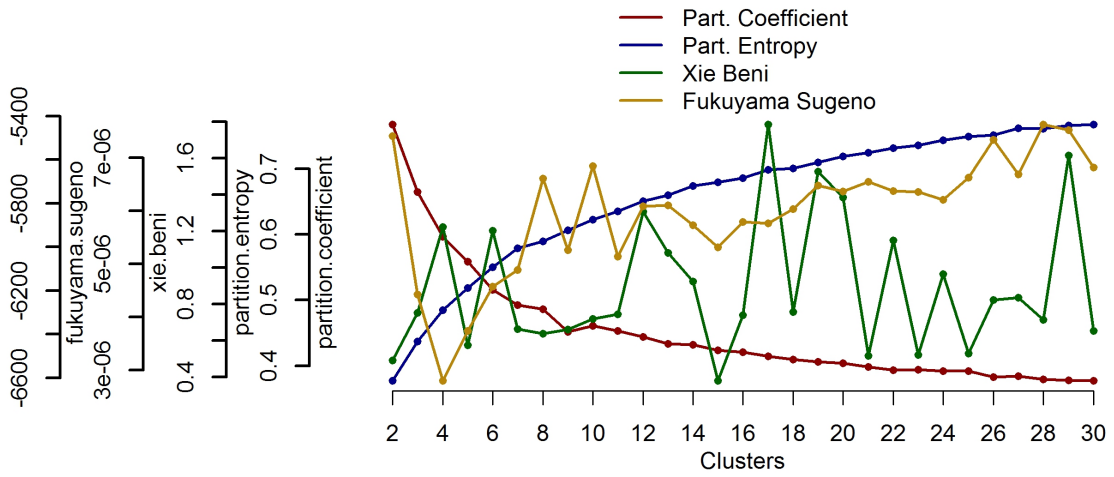


(b) $k = 9$



(c) $k = 13$

Figure 5.10 - Fuzzy C-Means cluster validation measures for different values of c for dataset S_0 and $m = 2$ considering three features.

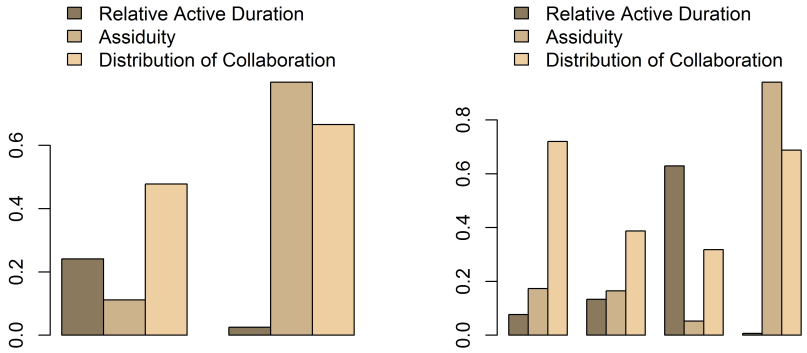


For the other datasets (S_1 to S_6) the same result could be observed. Moreover, different values of m were analyzed during the experiments in an attempt to get a clear evidence of clusters into data, but no evidence of cluster structure could be found.

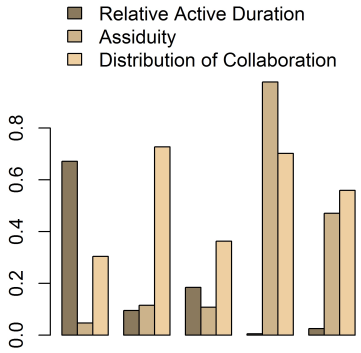
Figure 5.11 shows the centroids obtained through different values of c for dataset S_0. The values of c were chosen for the following reasons: $c = 2$ maximizes the value of Partition Coefficient and minimizes the value of Partition Entropy, $c = 4$ minimizes the value of Fukuyama Sugeno, $c = 15$ minimizes the value of Xie Beni and $c = 5$ visually seems to be the most appropriate value of c which maximizes Partition Coefficient and minimize the others. The groups observed for each value of c are similar to results obtained by K-means (compare Figure 5.11 to Figure 5.3 and 5.4). The increase of the partitions just shows slight variations in the values of features, which in practice doesn't change the meaning of the groups of volunteers already mentioned.

With regard to the datasets (S_0 to S_6) containing the eight features, S_1 and S_2 present the ideal condition where Partition Coefficient is maximum and the other is minimum for $c = 2$. Figure 5.12 shows the cluster validation measures for dataset S_1. The two groups obtained from S_1 and S_2 were similar.

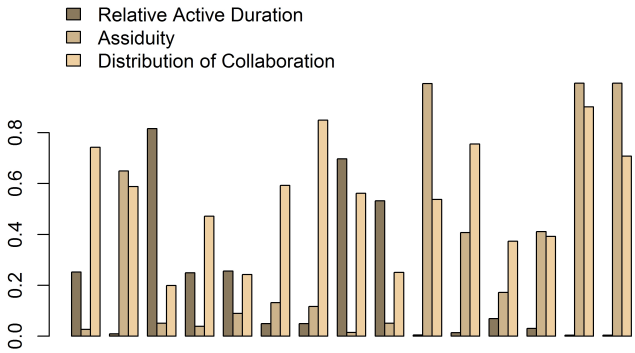
Figure 5.11 - The centroids from C-Means considering the dataset S_0 and $m = 2$.



(a) $c = 2$ (b) $c = 4$



(c) $c = 5$



(d) $c = 15$

Figure 5.13 shows the values of centroids. Observe that these groups are similar to the two groups found by K-Means (compare the Figure 5.13 to the Figure 5.7), the main particularity is the lower value to the feature Recurrence in Figure 5.13.

Figure 5.12 - Fuzzy C-Means cluster validation measures for different values of c for dataset S_1 and $m = 2$.

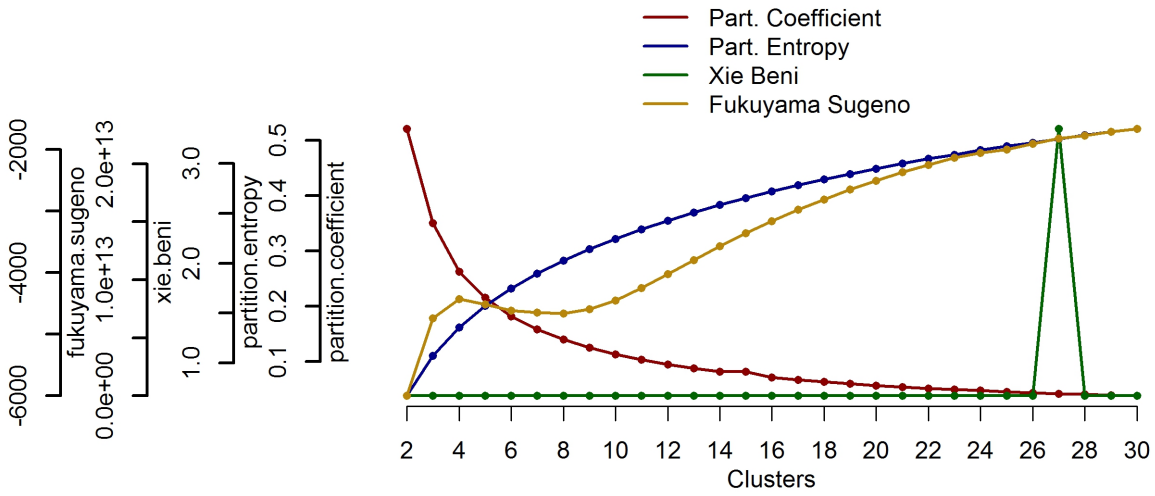
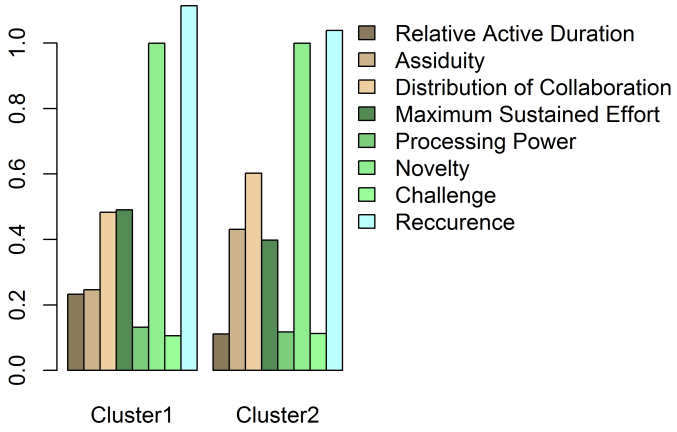


Figure 5.13 - The centroids from Fuzzy C-Means considering dataset S_1, $c = 2$ and $m = 2$ containing the eight features.



5.2.2 Density-based Method: DBSCAN

In contrast to partitioning methods, density-based approach does not require a number of partitions in which the data objects will be organized. Methods based on density locate regions of high density that are separated from one another by regions from low density (TAN et al., 2006). Regions of high density are considered clusters. The notion of density is estimated by counting the number of points within a specified radius (*Eps*) of a particular point, the region is said dense if the number of

points counted is higher than a given threshold ($MinPts$). In a density-based approach, the methods classify each point (i.e. each data object of the input dataset) as being in the interior of a dense region (a core point), on the edge of a dense region (a border point) or in a sparsely region (noise or outliers).

Popular density-based method are DBSCAN (ESTER et al., 1996) and its variations. The basic DBSCAN steps are described by the Algorithm 5.2.3. In short, the algorithm works in a way that any two core points within a distance of Eps of one another are put in the same cluster, while any border point that is close enough to a core point is put in the same cluster of this core point.

Algorithm 5.2.3: Basic DBSCAN algorithm

Input: A radius (Eps) and the minimum number of points ($MinPts$) required to define a dense region

Result: A set of clusters

```
1  ClusterId = 0;
2  for each point p in dataset do
3      if p is not visited then
4          mark p as visited;
5          count all points within Eps-neighborhood of p;
6          if Eps-neighborhood<MinPts then mark p as noise;
7          else
8              add p to ClusterId;
9              for each point p' into the neighborhood of p do
10                 if p is not visited then
11                     mark p' as visited;
12                     count all points within Eps-neighborhood of p';
13                     if Eps-neighborhood of p'>=MinPts then Add the
                        neighborhood of p' to p;
14                     if p' is not member of any cluster then add p' to ClusterId;
15                 end
16                 ClusterId++;
17             end
18         end
19     end
20 end
```

With regard to its strengths and weaknesses, DBSCAN algorithm is relatively resis-

tant to noise and outliers, being able to handle clusters of arbitrary shape and size. If compared to K-Means and its variations, this method can find many clusters that otherwise could not be found (TAN et al., 2006). Drawbacks of this method includes its limitations to detect clusters with widely varying densities, the issue to define density and the computation of nearest neighbors for high-dimensional data.

Although the number of clusters is not required, the choice of the parameters Eps and $MinPts$ may be not an easy task (AGGARWAL; REDDY, 2013). To determine these parameters Ester et al. (1996) proposed a heuristic called k-dist graph. The heuristic consists of computing the distance from a point to its Kth nearest neighbor, for all the data points. Sort them in increasing order and plot the sorted distance values. Based on this plot, the suitable value of Eps is the value where the plot changes sharply (i.e. "valley" of the graph) and the value of $MinPts$ is set to k. Note that the value of Eps depends on k, however, it does not change dramatically as k (i.e. $MinPts$) changes (TAN et al., 2006). On the other hand, if k is too small, noise or outliers will be incorrectly labeled as cluster, while for too large values of k small clusters are likely to be labeled as noise. For 2-dimensional data, the original DBSCAN algorithm used a value of $k = 4$, because it appears to be a reasonable value for most two-dimensional datasets (ESTER et al., 1996).

Table 5.4 shows the values of Eps suggested by the heuristic for different values of k ($MinPts$) and the respective number of clusters found by the algorithm, considering the dataset S_0 composed by the features Relative Active Duration, Assiduity and Distribution of Collaboration.

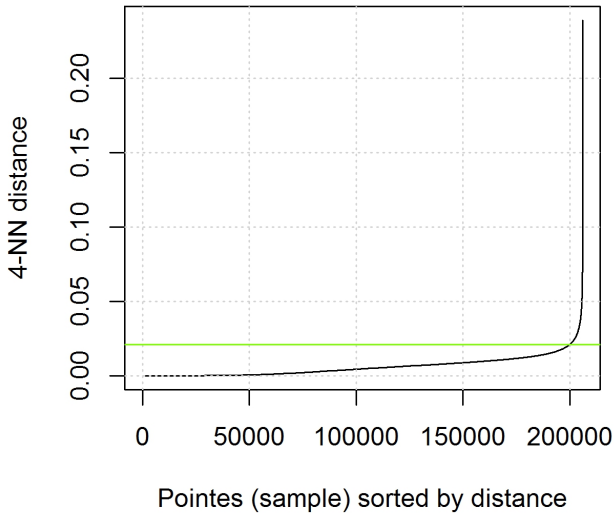
Table 5.4 - Values of Eps suggested by the heuristic for a given $MinPts$ and the number of clusters obtained by each set of parameters, considering the dataset S_0.

MinPts	Eps	Clusters
4	0.021	70
8	0.025	21
16	0.032	6
32	0.048	2
64	0.052	2
128	0.071	2
252	0.092	2
512	0.12	1

Figure 5.14 exemplifies how the value of $Eps = 0.021$ was chosen for $k = 4$ (i.e.

$MinPts$), the green horizontal line shows the "valley" of the 4-dist graph.

Figure 5.14 - 4-dist graph obtained through dataset S_0.



In all experiments conducted at least one of the clusters obtained from the seven datasets (S_0 to S_6), either composed by the three or the eight features, was composed by volunteers who joined the project just for few days and volunteers who joined the project for more than one year. In the context of extracting behavioral profiles, it is expected that these two kinds of volunteers are organized into different clusters. Moreover, it is notable that assiduous volunteers who joined the project for years and spread their collaboration along their active days were always classified as noise by DBSCAN algorithm.

5.2.3 Model-based Method: Self-Organizing Maps

Model-based methods are techniques that aim to optimize the fit between a given data set and some mathematical model. This category of methods is composed by different approaches, being the Neural Networks one of them. The Self-Organizing Maps (SOM) is likely the most widely used neural network algorithm. Proposed by Kohonen (1998) as a visualization tool, SOM has been used for data visualization and clustering. Details about the use of SOM in exploratory data analysis and cluster analysis are presented by Kaski (1997) and Vesanto et al. (2002).

The SOM algorithm implements a mapping of the input data space into a low-dimensional grid. The grid is usually one- or two- dimensional, mainly when the objective is data visualization. With regard to its shape, this grid may be defined as rectangular, hexagonal, or even irregular. Given a shape of the grid, each grid unit j has a prototype vector $m_j = m_{j1}, m_{j2}, ..., m_{jd}$ in a specific location r_j, where d represents the dimension of a data object.

The main characteristics of this mapping consist of compressing information through a set of prototype vectors. The prototype vectors (m_j) reproduce the original data as well as possible, while preserving the topology of the original data set, in a way that if two data samples are close to each other in the grid, they are likely to be close in the original data space. Formally, compress date and preserve the topology of the original data are called Quantization and Projection, respectively.

The basic SOM algorithm is shown in Algorithm 5.2.4. In short, the SOM algorithm is based on unsupervised learning which adjusts an initial grid with random prototype vectors into a grid that summarizes the data preserving their main characteristics. The grid adjusts to the data by adapting the values of its prototype vectors during a set of steps called learning. At each learning step t a sample $x_i = x_{i1}, x_{i2}, ..., x_{id}$ is chosen from input data, until all data are picked. Thus, the distances (usually the Euclidean distance) between x_i and all the prototype vectors are calculated to obtain the best-matching unit (BMU), i.e., the closest prototype vector. Once the closest prototype vector is found, its values and the values of its neighborhood prototype vectors are updated, moving them toward x_i. The update rule is described as:

$$m_i(t+1) = m_i(t) + \alpha(t) h_{ci}[x(t) - m_i(t)], \tag{5.3}$$

where $\alpha(t)$ is the learning rate at learning step t and $h_{c_j,j}$ is a neighborhood centered winner unit (BMU).

In literature both $\alpha(t)$ and the radius of neighborhood ($N_c(t)$) are usually decreasing monotonically in time (KOHONEN, 1997). For a SOM network with few hundred cells on grid, the selection of parameters is not very crucial, though special caution must be taken in the choice of initial radius of neighborhood. In relation to $\alpha(t)$, it can be linear, exponential, or inversely proportional to t. During the first 1000 steps when the ordering of the maps should take place, $\alpha(t)$ should be large (close to one) and $N_c(t)$ should shrink to perform fine-adjustment. After the ordering phase, $\alpha(t)$

Algorithm 5.2.4: Basic SOM algorithm

Input: Grid of reference vectors m_j, input data set x_i, initial learning rate $\alpha(0)$ and initial radius of neighborhood $N_c(0)$

Result: Mapping of the input data space into grid

```
Randomize the prototype vectors m_j;
while train do
    Draw exemplar x_i from the exemplar set;
    Find m_j with minimum distance from x_i;
    Update the prototype vectors into the neighborhood of m_j (Equation 5.3);
    Decrease α;
    Decrease radius N_c;
end
```

should attain small values (less than 0.02) over a long period. According to the "rule of thumb" proposed by Kohonen (1997) the number of steps must be at least 500 times the number of network units.

The grid obtained from SOM algorithm has been used for clustering analysis in at least for two different approaches. The first is called SOM-Based Clustering and consists of a 2-phase strategy. In this approach, SOM is trained and the resulting prototype vectors are clustered by standard clustering algorithms or by some adapted algorithm which take the SOM neighborhoods into account (VESANTO et al., 2002). More information about this approach is presented in (VESANTO; ALHONIEMI, 2000). The other is called Distance Matrix-based Clustering. This approach consists of assigning for each cell r_j a value which represents the distance between r_j and its neighbors on grid. The visual representation of the grid by paint, often in grayscale, each position r_j according to its values is usually called U-Matrix. Through this representation cluster may be detected by visual inspection or making use of some algorithm like proposed by Vesanto e Sulkava (2002). In short, groups of light colors are considered clusters, and the dark regions are the boundaries between clusters.

In order to investigate the Distance Matrix-based Clustering approach several experiments were done with the original dataset (S_0) and the datasets obtained through the standardization methods (S_1 to S_6). The experiments were organized into two parts: one considering just the features Relative Active Duration, Assiduity and Distribution of Collaboration and the other containing the entire set of proposed features. For each dataset, a set of U-Matrix was plotted and analyzed in order to detect groups of volunteers which follow a similar behavior, i.e., groups of light colors on U-Matrix. This set of U-Matrix was obtained as follows: given a dataset S_x

with three or eight features, several square grids varying from 10×10 to 40×40 by steps of 10 were obtained through the SOM algorithm; for each grid dimension $M \times M$, the SOM algorithm was run with different initial radius of neighborhood ($1 \leqslant N_c(0) \leqslant M$) providing a total of M U-Matrices by a given grid dimension.

With regard to the other parameters: the initial learning rate ($\alpha(0)$) was set as 1 and, the neighborhood h_{c_j} was defined as circular and both were decreased monotonically in time through the equations:

$$\alpha(t+1) = \alpha(t) * e^{-\frac{t}{300}} \tag{5.4}$$

$$N_c(t+1) = N_c(t) * e^{-\frac{t}{1000}} \tag{5.5}$$

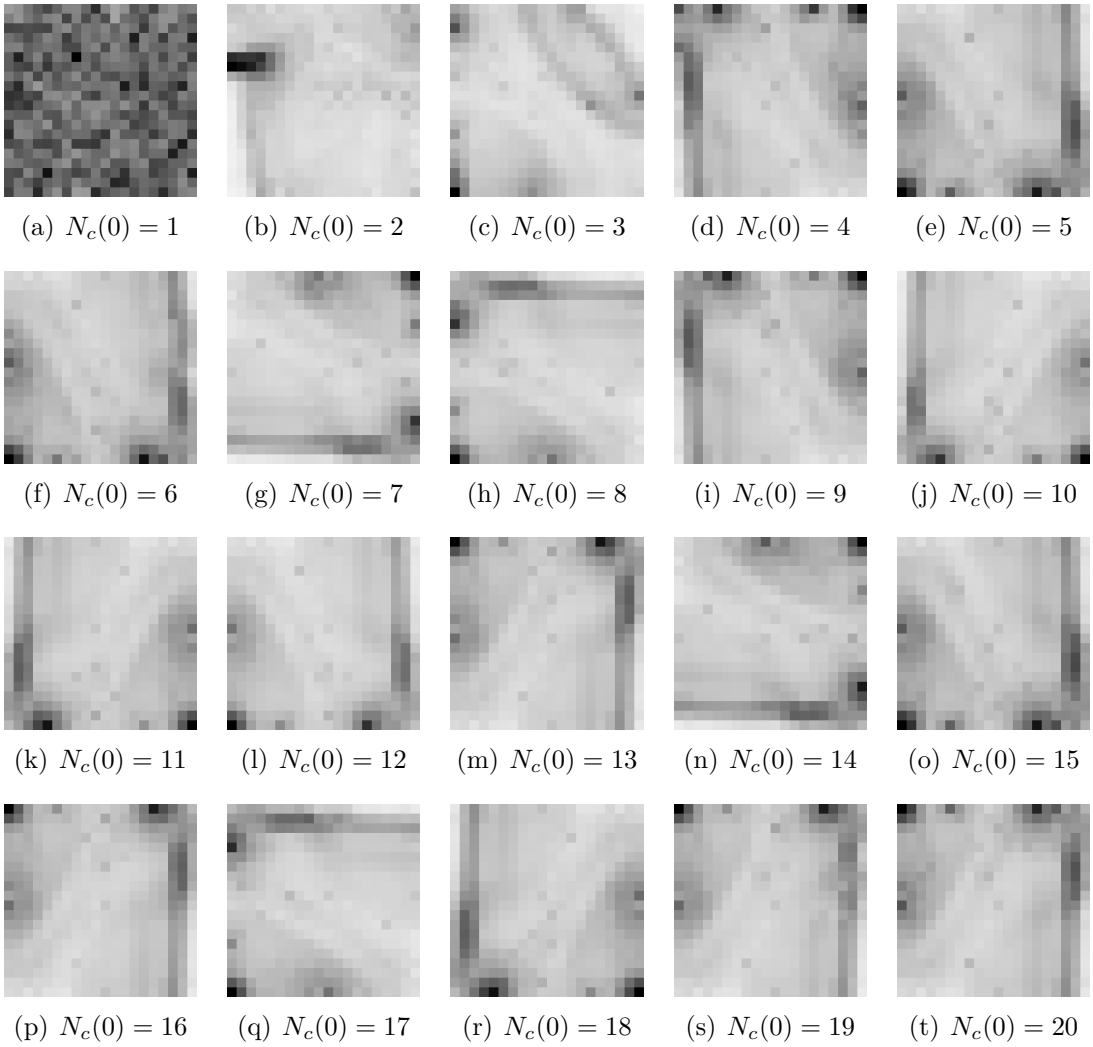
Figure 5.15 exemplifies the set of U-Matrix obtained through the experiments performed with dataset S_0 considering the features Relative Active Duration, Assiduity and Distribution of Collaboration for a grid of 20×20. Each figure (Figure 5.15(a) to Figure 5.15(t)) shows an U-Matrix obtained from the SOM algorithm for a given initial ratio.

Most of U-Matrices obtained on experiments were similar and did not suggest the evidence of clusters, i.e., we could not observe light regions delimited by borders in a dark gray scale. Figure 5.16 shows the U-Matrices where regions delimited by borders could be observed. These U-Matrices were obtained from experiments performed with datasets containing the features Relative Active Duration, Assiduity and Distribution of Collaboration. The analysis of these regions did not suggest new patterns of volunteers' behavior. With regard to the datasets containing all features proposed, it is important to note that the U-Matrices obtained on experiments did not present evidence of cluster.

5.2.4 Final Comments

Experiments showed that some volunteers categories may be described through clustering analysis. In general, the experiments showed that volunteers collaborated occasionally, regardless of the period of days elapsed between the first and last recorded collaboration. Assiduous volunteers who joined the project for years and spread their collaborations along their active days were not found during the experiments. It should be remembered that in Chapter 4, the exploratory data analysis

Figure 5.15 - U-Matrices obtained through the experiments performed with dataset S_0 considering the features Relative Active Duration, Assiduity and Distribution of Collaboration for a grid of 20×20.



(a) $N_c(0) = 1$ (b) $N_c(0) = 2$ (c) $N_c(0) = 3$ (d) $N_c(0) = 4$ (e) $N_c(0) = 5$

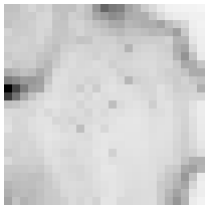
(f) $N_c(0) = 6$ (g) $N_c(0) = 7$ (h) $N_c(0) = 8$ (i) $N_c(0) = 9$ (j) $N_c(0) = 10$

(k) $N_c(0) = 11$ (l) $N_c(0) = 12$ (m) $N_c(0) = 13$ (n) $N_c(0) = 14$ (o) $N_c(0) = 15$

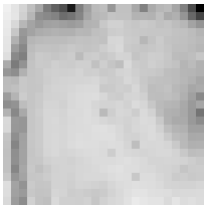
(p) $N_c(0) = 16$ (q) $N_c(0) = 17$ (r) $N_c(0) = 18$ (s) $N_c(0) = 19$ (t) $N_c(0) = 20$

(Section 4.2) singled out the evidence of this kind of volunteer behavior. As an alternative to apply standard clustering methods, the SOM algorithm may be used as a visualization tool to extract volunteers categories (MORAIS; SANTOS, 2015). On Chapter 6, this approach is presented in details.

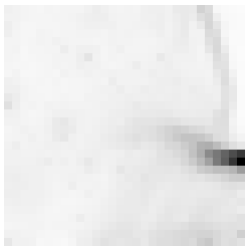
Figure 5.16 - U-Matrices obtained by different experiments where light regions are delimited by borders in a dark day scale.



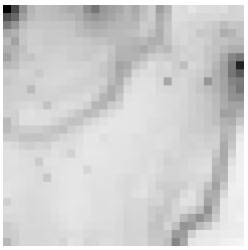
(a) S_4, $M = 25$ and $N_c(0) = 2$



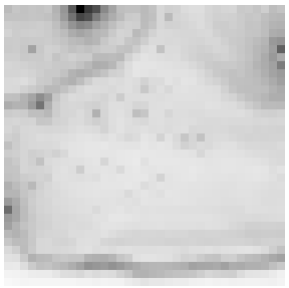
(b) S_5, $M = 25$ and $N_c(0) = 3$



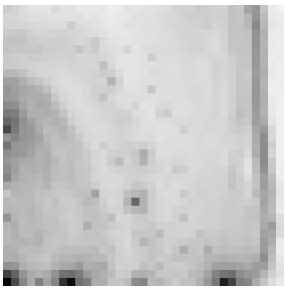
(c) S_6, $M = 30$ and $N_c(0) = 14$



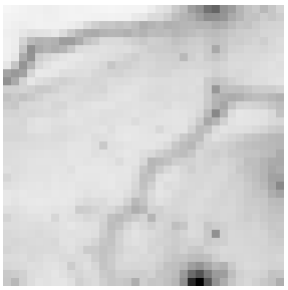
(d) S_3, $M = 30$ and $N_c(0) = 4$



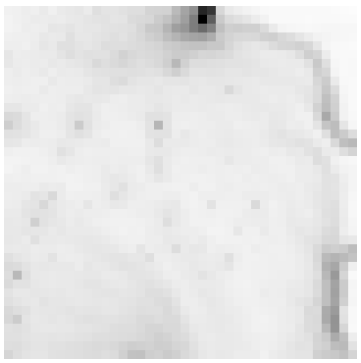
(e) S_1, $M = 35$ and $N_c(0) = 3$



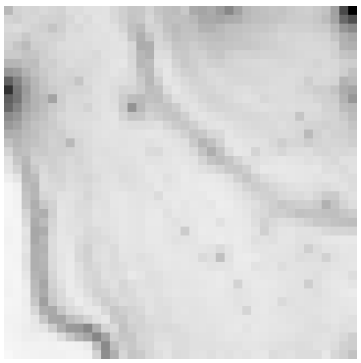
(f) S_2, $M = 35$ and $N_c(0) = 12$



(g) S_5, $M = 35$ and $N_c(0) = 2$



(h) S_0, $M = 40$ and $N_c(0) = 2$



(i) S_0, $M = 40$ and $N_c(0) = 4$

(j) S_3, $M = 40$ and $N_c(0) = 3$

6 DATA VISUALIZATION WITH SOM GRID

Some research works have taken advantage of the characteristic of compressing information and preservation of the topology of original data provided by SOM grid. A detailed study about the use of SOM in data analysis is presented by Vesanto (1999). In data analysis, SOM may act as a visualization tool. There are several approaches to use the SOM for data visualization. One of them consists of creating a graphical representation for each prototype vector and display them on the respective position (r_j) in the grid. In this approach, it is expected that the visual representation of grid holds interesting visual clues about the nature of the data.

Throughout this chapter we present how the combination of simple visual techniques backed up by a Self-Organizing Map may help to understand volunteers' interaction with web-based Citizen Science projects. This chapter is organized as follows: Section 6.1 describes a technique for visualization of multidimensional and/or large datasets that uses the Self-organizing Map as basis to cluster and reorganize data proposed in (MORAIS et al., 2014). Section 6.2 describes a way to use the grid SOM to visualize volunteers' behavior. Section 6.3 presents how the grid may aid the analysis of some additional characteristics (i.e. data) which were not used during the training steps of SOM algorithm. Section 6.4 summarizes the results achieved.

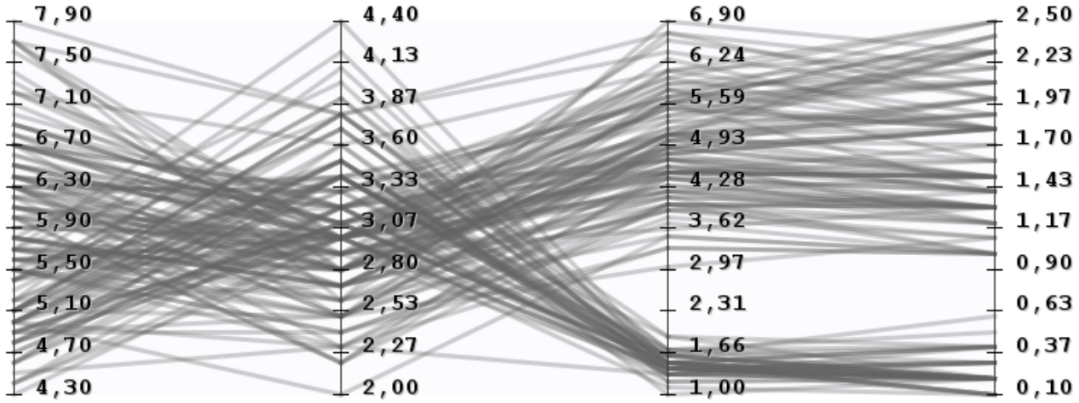
6.1 Data Visualization with a Self-Organizing Map and Parallel Coordinates

The grid produced by SOM algorithm can be considered a matrix, where each cell contains a prototypical data point m_{ij}, being i the row and j the column of matrix. The cells of this matrix can be used as a basis for well-established visualization techniques or new techniques more adequate to the problem and data. The choice of a visualization technique is largely intuitive and ad-hoc process which depends on the task being considered. Although there is no single rule for visual representation of usage logs processed by a SOM, Parallel Coordinates (INSELBERG, 2009) proved to be a good starting point (MORAIS; SANTOS, 2015).

Parallel Coordinates may be described as being a visual technique which maps a k-dimensional data to two-dimensional space by using k equidistant parallel axes. Each axis corresponds to one variable (dimension). In this technique one data object is presented as a line that crosses the axes according to the value of the variable for that axis. Figure 6.1 exemplifies this visual technique by showing the well-known Iris flower dataset. This dataset is formed by 50 samples of each Iris plant species:

Iris Setosa, Iris Virginica and Iris Versicolor. Each sample contains four variables: sepal length, sepal width, petal length and petal width. Iris flower dataset is known to have two easily separable clusters, being one formed by the samples of Iris Setosa and the other formed by the samples of Iris Virginica and Iris Versicolor.

Figure 6.1 - Iris data set visualized by Parallel Coordinates. From left to right each vertical line represents the variables: sepal length, sepal width, petal length and petal width.



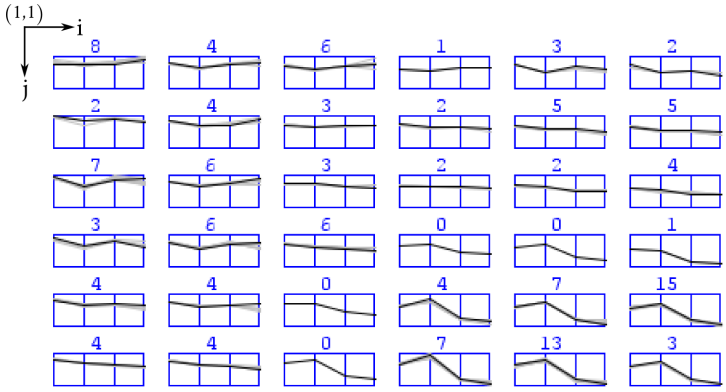
SOURCE: Morais et al. (2014)

Although the Iris dataset may be considered simpler dataset than the dataset used to describe volunteers' interaction, it is sufficient to demonstrate the concepts that involve the use of visual techniques backed up by a Self-Organizing Maps. To give an example of the technique proposed by Morais et al. (2014), in Figure 6.2 each prototype vector (m_{ij}) is displayed by dark polygons while the light ones represent the data items mapped by it. The number on top of each grid unit shows the amount of data represented by this grid SOM position. It is also important to point out that the component that displays a single parallel coordinate within the SOM was designed to be more concise, i.e. without labels and other information that could clutter the display composition itself.

In Figure 6.2 it is possible to observe a grouping of data with similar attributes in seven elements in the lower right corner (corresponding to the Iris Setosa samples, i.e. the group of flowers easily separable from the dataset). The other elements correspond to visual representations of Iris Virginica and Iris Versicolor samples and as it is previously known, it is hard to cluster these samples into two different groups. Observe that in Figure 6.2 the entire dataset is spread over the grid, it allows

a more efficient data analysis, once, in general, the clutter and data occlusion were reduced.

Figure 6.2 - Iris data set visualized by Parallel Coordinate over a SOM grid.



SOURCE: Morais et al. (2014)

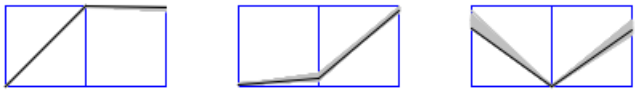
6.2 Visualizing Behavior with Parallel Coordinates and SOM

The combination of Parallel Coordinates backed up by a Self-Organizing Maps to visualize the features extracted from usage logs proved to be useful to help to understand volunteers' interaction with web-based Citizen Science projects as published in (MORAIS; SANTOS, 2015). To exemplify how to interpret the information held by this technique, we start by showing some examples of cells represented by Parallel Coordinates considering the features: Relative Active Duration, Assiduity and Distribution of Collaboration.

Figure 6.3 shows some cells chosen from a 15×15 grid which was trained with dataset S_0. In Figure 6.3 the prototypical data point (i.e. m_{ij}) is represented in cell as a black line, while the data mapped to the cell is represented in the background as gray lines. In this technique, each cell (i, j) can be viewed as a cluster, where the prototypical data point m_{ij} is the centroid of the original data mapped to the cell (i, j) (MORAIS et al., 2014).

Consider each axis of Parallel Coordinates represented by (blue) vertical lines and organized, from left to right, in the following order: (1) Relative Active Duration, (2) Assiduity and (3) Distribution of Collaboration. The first axis indicates the life cycle of a volunteer, i.e., short values represent that the volunteer joins and leaves the project soon thereafter, while higher values represent that the volunteer

Figure 6.3 - Visual representation of the cells of a 15×15 grid.



(a) Cell $(2, 15)$ (b) Cell $(15, 15)$ (c) Cell $(15, 7)$

(d) Cell $(1, 3)$ (e) Cell $(15, 1)$

joins the project for a while. The second axis measures how active the volunteer was during his/her life cycle, i.e., values close to one indicate that the volunteer collaborated almost every day with the project, while values close to one indicate that the volunteer had more inactive days than active days. Finally, the third axis measures how the collaborations were distributed during active days, thus values close to zero indicate that the collaborations were well distributed while values close to one shows that almost all collaborations where done in one day. Therefore, the cells exemplified in Figure 6.3 can be interpreted as:

a) Cell $(2, 15)$: represents volunteers who joined the project and abandoned shortly afterward, were assiduous and have performed almost all collaboration in one day;

b) Cell $(15, 15)$: represents volunteers who joined the project by some months, were not assiduous and have performed almost all collaboration in one day;

c) Cell $(15, 7)$: represents volunteers who joined the project for almost one year, were not assiduous and have performed almost all collaboration in one day;

d) Cell $(1, 3)$: represents volunteers who joined the project and abandoned it shortly afterward, but in contrast to other examples, volunteers mapped by this cell were not assiduous and distributed their collaborations along the active days;

e) Cell $(15, 1)$: represents volunteers who joined the project for more than one year, were assiduous and distributed their collaboration along the active days.

Figure 6.4 shows the 15×15 grid where the examples of cells shown in Figure 6.3 were extracted. Coordinates of the cells were enumerated based on grid dimension, from left to right and from top to bottom, starting with one. Moreover, note that in the top right of each cell the number of original data mapped to respective position is shown. This visual representation sheds light on the behavior of all volunteers. Through this technique different visual patterns leap to the eye, allowing that the analyst finds isolated patterns or regions with similar patterns. Each visual pattern can be interpreted as a volunteer's profile. However, it should be noted that the concept of volunteers' profiles is subjective and depends on the interpretation of the analyst.

Figure 6.4 - Visualization of dataset S_3 (containing three features) through Parallel Coordinates and SOM grid.

In Figure 6.4 it is possible to identify the patterns obtained during the experiments performed on Chapter 5 and some new patterns. The patterns described by the cells $(2, 15)$, $(15, 15)$ and $(15, 7)$ (and their neighborhood) were detected by experiments

performed with K-Means and Fuzzy C-Means, while the cells $(1, 3)$, $(15, 1)$ are examples of patterns that were not detectable by the standard clustering methods analyzed in Chapter 5. The patterns described by these cells were found for all grid trained with initial radius larger than two, regardless of the dataset (S_0 to S_6).

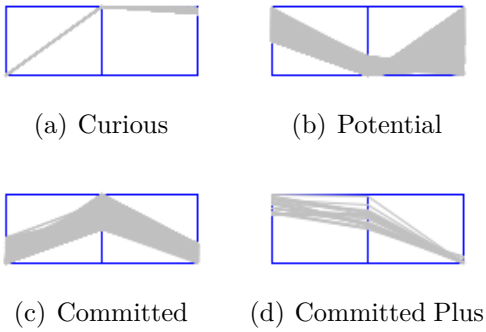
Based on features Relative Active Duration, Assiduity and Distribution of Collaboration, some of the patterns may be labeled (MORAIS; SANTOS, 2015) as:

- **Curious:** volunteers who have high assiduity, but joined the project and abandoned it shortly afterward (low relative activity duration) doing all collaborations in the few days of participation (high distribution of collaboration);
- **Potential:** volunteers that could be motivated to collaborate more. These volunteers have medium to high relative activity duration and below average assiduity, indicating that for some reason those volunteers were attracted by the project, collaborated with it for a while, left and then returned again after some time;
- **Committed:** volunteers who have medium to high assiduity and below average distribution of collaboration;
- **Committed Plus:** corresponding to committed volunteers with a high relative activity duration, in other words, committed volunteers who joined the project for more than one year.

Figure 6.5 shows the visual patterns labeled by Morais e Santos (2015). Observe that more than one polygon were drawn in each sub-figure. The set of polygons represented in each sub-figure are slight variations of a visual pattern which may be labeled.

Figure 6.6 shows the graphic representation of dataset S_3 containing the eight features through a 15×15 grid. In Figure 6.6 some cells had their background painted according to the labels proposed by Morais e Santos (2015). The axes of Parallel Coordinates are organized, from left to right, in the following order: (1) Relative Activity Duration, (2) Assiduity and (3)Distribution of Collaboration, (4) Maximum Sustained Effort, (5) Processing Power, (6) Novelty, (7) Challenge and (8) Recurrence. Through Figure 6.6 it is possible to note that in contrast to the first three features, the last five features have the same values in almost all prototypes (i.e. cells).

Figure 6.5 - Examples of visual patterns labeled by Morais e Santos (2015).



(a) Curious (b) Potential

(c) Committed (d) Committed Plus

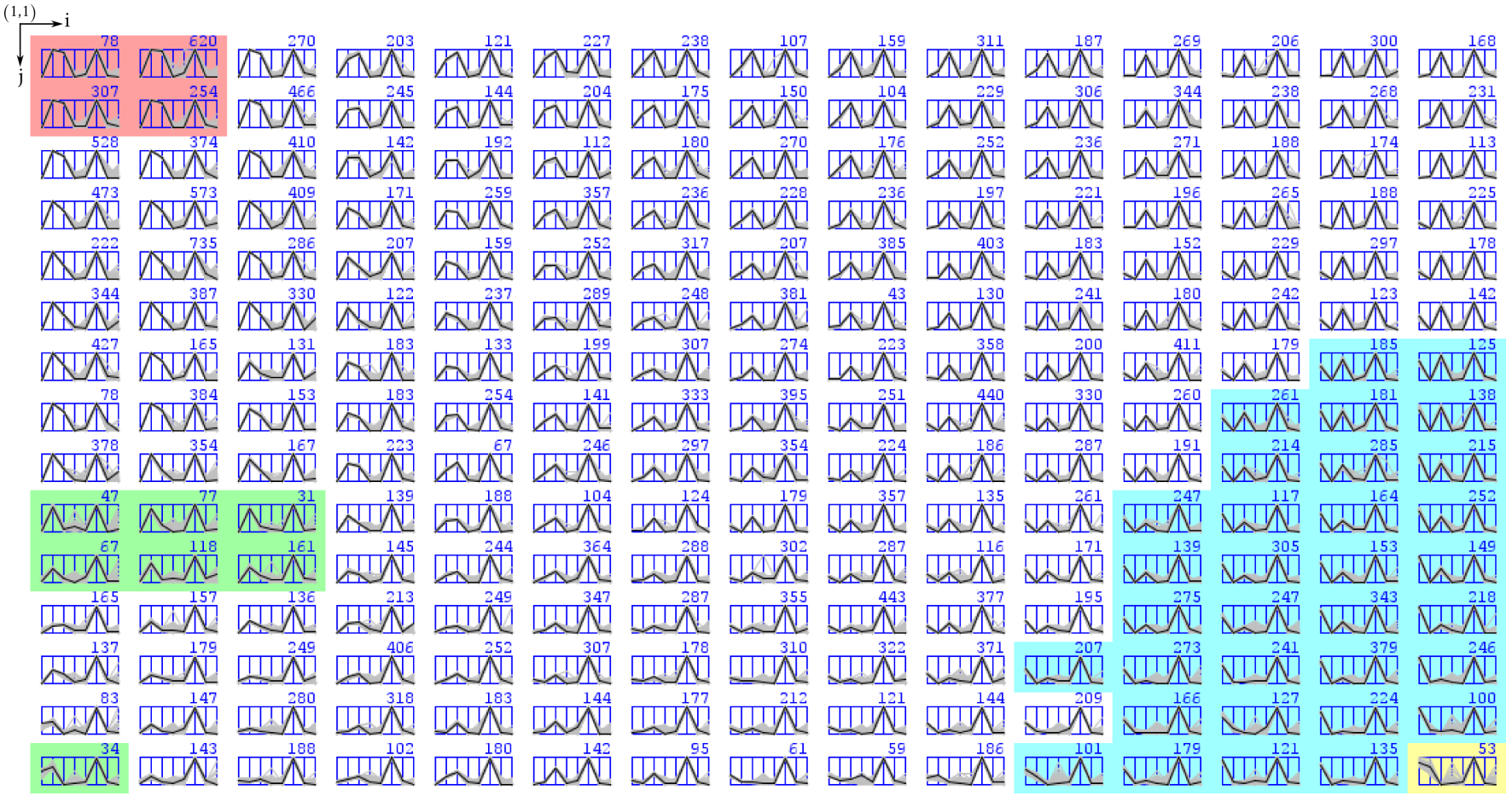
Analysis of the visual patterns shows that most of volunteers may be characterized by maintaining his/her mental context for few minutes (low value on fourth axis), performing few collaborations per second (low value on fifth axis), usually having to deal with easy task (low value on seventh axis) and had never returned in the same active day to collaborate (low value on eighth axis). Some exceptions may be observed, for example, the cell $(15, 15)$ in Figure 6.6 shows that there are volunteers who maintained his/her mental context for longer periods. These exceptions appeared on different visual patterns (i.e. position of grid) like in cells $(14, 1)$, (10,9) and others. Hence, this characteristic seems not to be associated with a particular kind of volunteer's profile.

6.3 Visualizing Other Features

Besides helping the visual detection of volunteers' profiles, the grid may also aid in the analysis of some additional characteristics through the distribution of the data over the grid. The coordinates of the grid provide a link between the features which define a volunteer's behavior and any other data about this volunteer. Each position (i,j) of grid can be used to show a summary of the statistics of the (additional) data. Figure 6.7 shows this approach in order to evaluate the quality of collaboration of volunteers. In each grid position, we plotted the histogram of feature agreement of the volunteers mapped to this position. In order to be more concise, each histogram was plotted without labels and other information that could clutter the display composition itself.

It is also important to point out that Figure 6.7 consists of another visual representation of the same 15×15 grid used to plot the Figure 6.4. Some grid positions were highlighted in different colors according to patterns labeled by Morais e Santos

Figure 6.6 - Visualization of dataset S_3 (containing eight features) through Parallel Coordinates and SOM grid.



Grid positions were highlighted in different colors according to profiles labeled by Morais e Santos (2015) red for “curious”, blue for “potential”, green for “committed” and yellow for “committed plus”.

(2015): red for “curious”, blue for “potential”, green for “committed” and yellow for “committed plus”. Through Figure 6.7 we show that the feature agreement presents a positive distribution in all grid positions. It means that volunteers usually have a good quality of their collaboration (more than 75% accuracy) regardless of their kind of behavior.

Figure 6.7 - Distribution of feature agreement over grid.

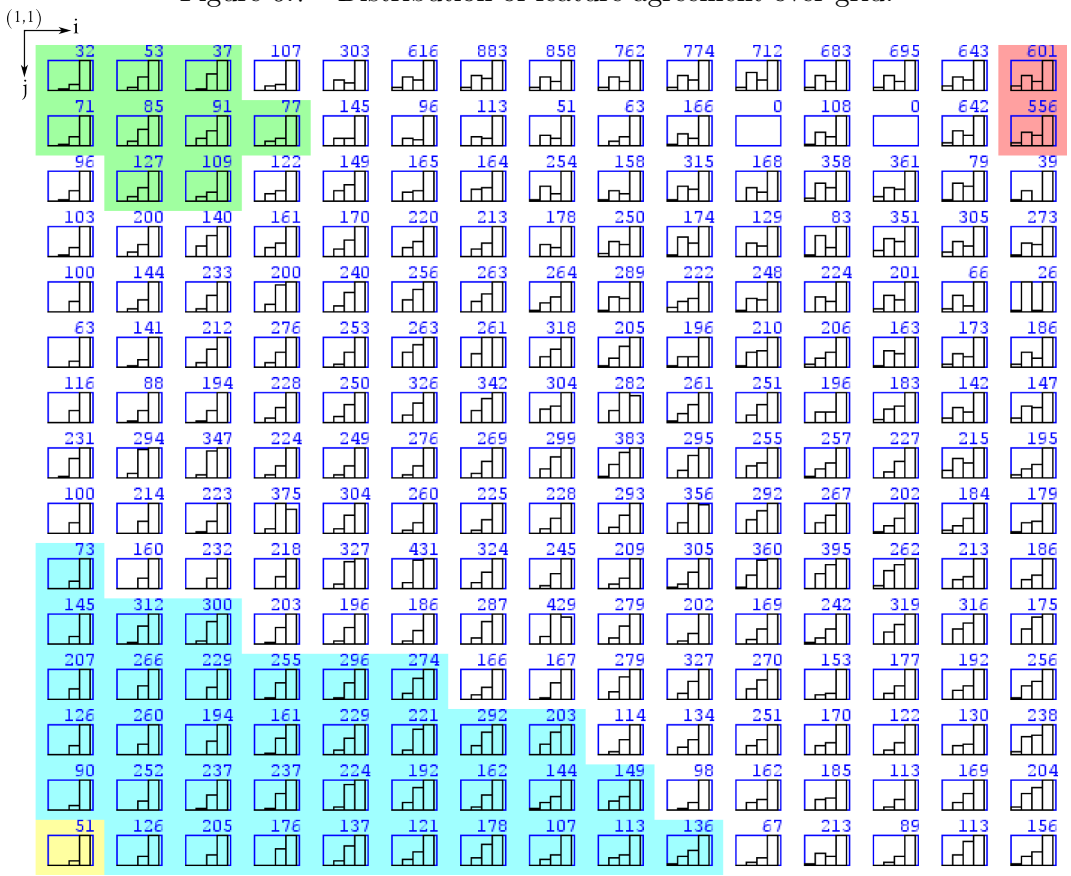
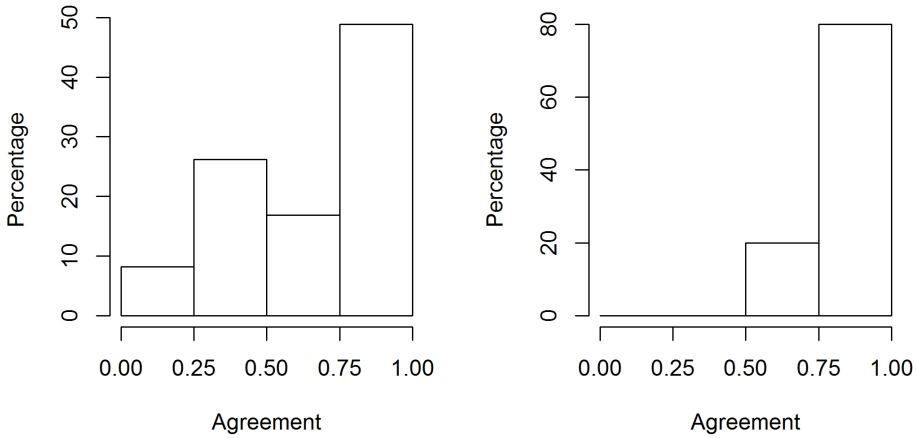


Figure 6.7 also suggests that occasional volunteers may be more likely to have low grades compared to volunteers who collaborated through a committed involvement. To clarify this observation consider the complete histogram of cells $(1, 15)$ and $(15, 1)$ shown in Figure 6.8. Figure 6.8(a) shows the existence of volunteers with lower collaboration quality (less than 40% accuracy) between the volunteers labeled as curious, which is not observed in volunteers labeled as committed plus, see Figure 6.8(b).

Figure 6.8 - Complete histogram of cells $(1, 15)$ and $(15, 1)$.



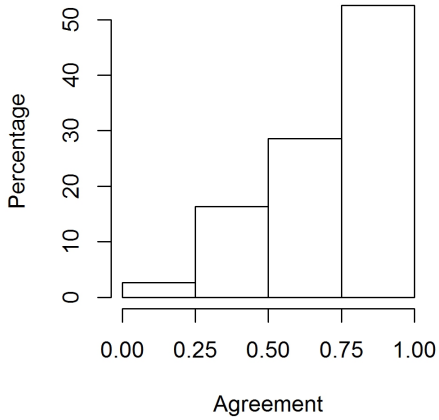
(a) Cell $(1, 15)$ (b) Cell $(15, 1)$

6.4 Final Comments

The features extracted from our case study resulted on a dataset with an almost homogeneous distribution. Although there are no clear evidences of clusters into data, the combination of simple visual techniques backed up by a Self-Organizing Maps (SOM) proved to be useful to point out volunteers' categories. With regard to the existence of behavioral profiles with higher collaboration quality compared to others, this chapter indicates that although some occasional volunteers might be less motivated if compared to volunteers who collaborated through a committed involvement, in general, all volunteers (regardless of their kind of behavior) have a good quality on their collaboration (more than 75% accuracy).

The volunteers removed from data (the ones who collaborate just for one day) were labeled as curious in (MORAIS; SANTOS, 2015). The analysis of these volunteers shows that most of them are characterized by maintaining his/her mental context for few minutes, performing few collaborations per second, usually having to deal with easy task and had never returned in the same active day to collaborate, with some exceptions of volunteers who returned at least two times in the same active day, see Figure A.2 in appendix. Figure 6.9 shows that the accuracy of these volunteers is similar to volunteers labeled as curious and joined the project for more than one day.

Figure 6.9 - Distribution of accuracy of volunteers who collaborate just one day.



7 CONCLUSIONS

This research was set out to extract behavioral profiles from records collected by web-based citizen science projects. At least, web-based citizen science projects collect which volunteer (anonymous IDs) did what (the volunteer's collaboration) and when (timestamps of registered collaboration). As far as we know, to date, little work have taken advantage from these records in order to use them as a feedback tool about volunteers' motivation. These kinds of records (called usage logs by this document) demonstrated to be a potential source of information.

In section 4.1 relatively simple visual techniques shed light on what kind of information may be hidden on usage logs. Through these techniques we were able to get some insights into volunteers' general behaviors during the project's duration. These visual techniques proved be useful to highlight information regarding the results of polices to get volunteers, actions to keep them collaborating and volunteers' abandonment.

In order to better understand the details about how volunteers spend their time interacting with web-based citizen science projects, we rewrote a model for human interaction with technology proposed by O'Brien e Toms (2008) through the definition of work sessions suggested by Mao et al. (2013). Based on this, we proposed eight measures of interaction (features) with web-based citizen science projects for volunteers' characterization.

In a first analysis, each feature proved to be able to describe some behavioral aspect of the volunteers. The first insights on data distributions indicated how challenging can be to apply clustering algorithms in order to find volunteers who exhibit similar behavior. The features extracted from usage logs of our study case, may be initially described as a homogeneous data.

We investigated two approaches with the goal of finding evidences of groups on data. First, we equalized the size and variability of our data through six different standardization methods, thus we observed the projection of data obtained through these methods. Two of them suggest at least two regions with higher density of points. In the second approach, we performed some experiments in other to evaluate the relevance of each feature on grouping data. Through these experiments we noted that the features Relative Activity Duration, Assiduity and Distribution of Collaborations seemed to be the most relevant features to point out groups in our study case.

To explore the use of clustering methods on the detection of volunteers' profile, we investigated methods with different clustering methodology. They were two partitioning methods (K-Means and Fuzzy C-Means), a density-based method (DBSCAN) and a model-based method (Self-Organizing Map). Experiments showed that some volunteers profiles may be found through K-Means, Fuzzy C-Means and Self-Organizing Map. However, it should be noted that none of them was able to detect profiles which describe the assiduous volunteers who joined the project for years. As an alternative to apply clustering methods, this work proposed a combination of visual techniques backed up by a Self-Organizing Maps (SOM). It organized each volunteer into a bi-dimensional grid. This representation allows that different visual patterns draw attention. This approach proved to be useful to support visual identification of patterns of behavior and outliers. Some of the patterns found can be labeled as curious, potential, committed and committed plus.

Through this grid we assessed the existence of behavioral profiles with higher collaboration quality compared to others. It demonstrated how these approaches may be used in order to visualize other volunteers' features which were not included in the initial construction of the grid. These experiment showed that volunteers usually have a good quality in their collaboration (more than 80% accuracy) regardless of their behavior.

Throughout this chapter, the main contributions are highlighted in section 7.1 and perspectives of future research are presented in section 7.3.

7.1 Contributions

The main contributions of this research were:

a) We rewrote, in Section 3.3.1, a model for human interaction with technology in order to set it in the context of web-based citizen science projects. This model provides some insights about what kind of information should be extracted from usage logs to describe volunteers, while the feature proposed shed light on a way to extract them from usage logs. We expect that it may inspire the extraction of new features;

b) We described, in Section 4.1, two visual techniques which were able to provide insights into volunteers' general behaviors during the project's duration;

c) We built, in Section 4.2, a dataset containing information about the fea-

tures extracted from usage logs of almost 150.000 volunteers. We expect that it can be used in other works;

d) We analyzed, in Chapter 5, some standard clustering techniques emphasizing the challenges and issues of applying clustering techniques to detect groups of volunteers with similar pattern behavior;

e) We presented, in Chapter 6, a technique for data visualization comprised of traditional visual components organized in a SOM grid. This technique allowed the visualization of volunteers' characteristic and proved to be useful to group the dataset, making easier the identification of common patterns and outliers. Code and examples will be available to download as soon as possible.

7.2 Publications

During the development of this research work the following papers were published:

a) *Icon and Geometric Data Visualization with a Self-Organizing Map Grid* (MORAIS et al., 2014): In this paper we proposed a combination of icon-based and geometric-based visualization techniques backed up by a Self-Organizing Map (SOM). This approach allows dimensionality reduction and topology preservation which may aid to visualize data, reducing clutter and facilitating identification of associations, clusters and outliers. This paper presents in details the concepts used in Chapter 6 and emphasizes some theoretical aspects of using a Self-Organizing Map as a visualization tool;

b) *Neural network based visualization of collaborations in a citizen science project* (MORAIS et al., 2014): This paper presents the first insights of how the use of SOM as a visualization tool may be helpful to point out behavioral aspects of volunteers. In this publication we used the measures of interaction proposed in (MORAIS et al., 2013), once we had not rewritten the a model for human interaction with technology;

c) *Visualization of Citizen Science Volunteers' Behaviors with Data from Usage Logs* (MORAIS; SANTOS, 2015): Using the measures of interaction Relative Activity Duration, Assiduity and Distribution of Collaborations, this

paper presents how visualization techniques may help infer volunteer behavior from usage logs. Through this paper we showed that some behavioral aspects may be detect from these three measures of interaction.

d) *Challenges in mapping behaviors to activities using logs from a citizen science project* (MORAIS et al., 2016): In this paper we published the measures of interaction proposed in Chapter 3, section 3.3.1. We also commented on the applicability of those measures and described an approach which may yield more precise logs, i.e. kind of data that may be registered by web-based citizen science projects which may help shed light on some of the engagement attributes described by O'Brien and Toms (O'BRIEN; TOMS, 2008) that could not be measure from registers of who did what and when. Through this work we recognized the necessity of define the pipeline, described in Section 3.3, used to scrap, filter and format the raw data.

7.3 Future Work

To help future investigations and allow that other studies apply the methodology used by this work, in short term, we are implementing our codes in R - a free software environment for statistical computing and graphics which is widely used in the data science context - and making it available through the distribution of R packages.

As a continuation of this work, the following issues may be considered:

a) Future investigations may be done in order to assess whether the behavioral aspects found by this research are common characteristics in web-based citizen science projects;

b) Future works may focus on the proposal of methodologies to detect temporal variation on volunteers' behaviors;

c) Web-based citizen science projects are able to collect more detailed data than registers of who, did what and when (usage logs). Frameworks like USABILICS (VASCONCELOS; BALDOCHI JR., 2012) allow the collection of data from the volunteer's browser, providing detailed information on interaction with the web interface. Examples of such data are: mouse movement, scroll bar usage, window resizing, graphs of visited pages, page loading times and interaction with web components such as links, forms, buttons, etc. As far as we know, up to date, frameworks like USABILICS have not being used

in any web-based citizen science project. Future works may focus in collect such data and investigate what kind of information can be extracted from these records to better describe the volunteers' interaction with web-based citizen science projects.

REFERENCES

AGGARWAL, C. C.; REDDY, C. K. **Data clustering**: algorithms and applications. [S.l.]: CRC Press, 2013. ISBN 978-1-4665-5822-9. 47, 50, 51, 52, 56, 69

AGRAWAL, R.; GEHRKE, J.; GUNOPULOS, D.; RAGHAVAN, P. Automatic subspace clustering of high dimensional data for data mining applications. In: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 1998, Seattle, Washington, USA. **Proceedings...** New York, NY, USA: ACM, 1998. p. 94–105. ISBN 0-89791-995-5. 52

ALABRI, A.; HUNTER, J. Enhancing the quality and trust of citizen science data. In: INTERNATIONAL CONFERENCE ON E-SCIENCE, 6., 2010, Brisbane, QLD. **Proceedings...** [S.l.]: IEEE, 2010. p. 81–88. ISBN 978-1-4244-8957-2. 7

ANKERST, M.; BREUNIG, M. M.; KRIEGEL, H.-P.; SANDER, J. Optics: ordering points to identify the clustering structure. In: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 1999, Philadelphia, Pennsylvania, USA. **Proceedings...** New York, NY, USA: ACM, 1999. p. 49–60. ISBN 1-58113-084-8. 51

ARCANJO, J. S.; LUZ, E. F.; FAZENDA, A. L.; RAMOS, F. M. Methods for evaluating volunteers' contributions in a deforestation detection citizen science project. **Future Generation Computer Systems**, Elsevier, v. 56, p. 550–557, 2016. 12

ARCANJO, J. S.; LUZ, E. F. D.; FAZENDA, A. L.; RAMOS, F. M. Evaluating volunteers' contributions in a citizen science project. In: INTERNATIONAL CONFERENCE ON E-SCIENCE, 10., 2014, Sao Paulo. **Proceedings...** [S.l.]: IEEE, 2014. p. 21–28. ISBN 978-1-4799-4288-6. 12

BADIE, B.; BERG-SCHLOSSER, D.; MORLINO, L. **International encyclopedia of political science**. [S.l.]: Sage, 2011. 21

BONNEY, R.; COOPER, C. B.; DICKINSON, J.; KELLING, S.; PHILLIPS, T.; ROSENBERG, K. V.; SHIRK, J. Citizen science: a developing tool for expanding science knowledge and scientific literacy. **BioScience**, Oxford University Press, v. 59, n. 11, p. 977–984, 2009. 5, 6, 8

BRABHAM, D. C. **Crowdsourcing**. [S.l.]: Mit Press, 2013. 17

BROOKING, C.; HUNTER, J. Reputation-aware filtering services for citizen science data. In: INTERNATIONAL CONFERENCE ON E-SCIENCE, 2011, Stockholm. **Proceedings...** [S.l.]: IEEE, 2011. p. 7–13. ISBN 978-1-4673-0026-1. 5, 7

CHI, Z.; YAN, H.; PHAM, T. **Fuzzy algorithms**: with applications to image processing and pattern recognition. [S.l.]: World Scientific, 1996. 51, 55, 57

CLEVELAND, W. S. et al. **The elements of graphing data**. [S.l.]: Wadsworth Advanced Books and Software Monterey, CA, 1985. 36

COHN, J. P. Citizen science: Can volunteers do real research? **BioScience**, Oxford University Press, v. 58, n. 3, p. 192–197, 2008. 1, 5, 8

DARCH, P. Managing the public to manage data: Citizen science and astronomy. **International Journal of Digital Curation**, v. 9, n. 1, p. 25–40, 2014. 8

DROEGE, S. Just because you paid them doesnt mean their data are better. In: CITIZEN SCIENCE TOOLKIT CONFERENCE, 2007, Ithaca, NY. **Proceedings...** Ithaca, New York, USA: Cornell Laboratory of Ornithology, 2007. p. 13–26. 5

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2., 1996, Portland, Oregon, USA. **Proceedings...** Menlo Park, CA: AAAI Press, 1996. p. 226–231. ISBN 1-57735-004-9. 51, 68, 69

EVELEIGH, A.; JENNETT, C.; BLANDFORD, A.; BROHAN, P.; COX, A. L. Designing for dabblers and deterring drop-outs in citizen science. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 2007, Toronto, Ontario, Canada. **Proceedings...** New York, NY, USA: ACM, 2014. p. 2985–2994. ISBN 978-1-4503-2473-1. 8, 12, 53

EVERITT, B.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster Analysis**. [S.l.]: Chichester, UK: Wiley, 2011. 36, 37, 47, 48

GAN, G.; MA, C.; WU, J. **Data clustering**: theory, algorithms, and applications. [S.l.]: Siam, 2007. xix, 39, 40

GOODCHILD, M. F. Citizens as sensors: the world of volunteered geography. **GeoJournal**, Springer, v. 69, n. 4, p. 211–221, 2007. 7

GRAY, J.; LIU, D. T.; NIETO-SANTISTEBAN, M.; SZALAY, A.; DEWITT, D. J.; HEBER, G. Scientific data management in the coming decade. **ACM SIGMOD Record**, ACM, v. 34, n. 4, p. 34–41, 2005. 1, 5

HAN, J.; KAMBER, M. **Data mining**: concepts and techniques. [S.l.]: Morgan kaufmann, 2006. 48, 50, 51, 52, 54

HINNEBURG, A.; KEIM, D. A. An efficient approach to clustering in large multimedia databases with noise. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 4., 1998, New York, New York. **Proceedings...** Menlo Park, CA: AAAI Press, 1998. p. 58–65. ISBN 978-1-57735-070-5. 51

HUSSON, F.; LÊ, S.; PAGÈS, J. **Exploratory multivariate analysis by example using R**. [S.l.]: CRC press, 2010. 38

INSELBERG, A. **Parallel coordinates**. [S.l.]: Springer, 2009. 77

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. [S.l.]: Prentice-Hall, Inc., 1988. 52

JOLLIFFE, I. **Principal component analysis**. [S.l.]: Wiley Online Library, 2002. 37, 38

KASKI, S. **Data exploration using self-organizing maps**. Tese (Doutorado) — PhD thesis, Helsinki University of Technology, Espoo (1997-03-21 1997), 1997. 70

KEIM, D. A.; KRIEGEL, H.-P. **Issues in visualizing large databases**. [S.l.]: Springer, 1995. 1

KOHONEN. The self organizing map. **Neurocomputing**, Elsevier, v. 21, n. 1, p. 1–6, 1998. 70

KOHONEN, T. **Self-organizing maps, 2nd Edition**. [S.l.]: Springer Verlag, 1997. 71, 72

LINTOTT, C.; SCHAWINSKI, K.; BAMFORD, S.; SLOSAR, A.; LAND, K.; THOMAS, D.; EDMONDSON, E.; MASTERS, K.; NICHOL, R. C.; RADDICK, M. J. et al. Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. **Monthly Notices of the Royal Astronomical Society**, Oxford University Press, v. 410, n. 1, p. 166–178, 2011. 6

LINTOTT, C. J.; SCHAWINSKI, K.; SLOSAR, A.; LAND, K.; BAMFORD, S.; THOMAS, D.; RADDICK, M. J.; NICHOL, R. C.; SZALAY, A.; ANDREESCU, D. et al. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. **Monthly Notices of the Royal Astronomical Society**, Oxford University Press, v. 389, n. 3, p. 1179–1189, 2008. 6, 8, 14, 20, 27, 34

MAO, A.; KAMAR, E.; HORVITZ, E. Why stop now? predicting worker engagement in online crowdsourcing. In: CONFERENCE ON HUMAN COMPUTATION AND CROWDSOURCING, 1., 2013, Palm Springs, California. **Proceedings...** Palo Alto, California: The AAAI Press, 2013. ISBN 978-1-57735-607-3. 2, 9, 12, 16, 17, 18, 53, 89

MAZZA, R. **Introduction to information visualization**. [S.l.]: Springer Science & Business Media, 2009. 23

MILLIGAN, G. W.; COOPER, M. C. A study of standardization of variables in cluster analysis. **Journal of classification**, Springer, v. 5, n. 2, p. 181–204, 1988. 38, 39, 40

MORAIS, A. M.; RADDICK, J.; SANTOS, R. D. C. dos. Visualization and characterization of users in a citizen science project. In: NEXT-GENERATION ANALYST, 2013, Baltimore, Maryland. **Proceedings...** [S.l.]: SPIE, 2013. ISBN 9780819495495. xix, 2, 9, 10, 11, 19, 23, 24, 25, 26, 53, 91

MORAIS, A. M.; SANTOS, R. D.; RADDICK, M. J. Neural network based visualization of collaborations in a citizen science project. In: NEXT-GENERATION ANALYST II, 2014, Baltimore, Maryland. **Proceedings...** [S.l.]: SPIE, 2014. p. 912207–912207. ISBN 9781628410594. 91

MORAIS, A. M.; VASCONCELOS, L. G. de; SANTOS, R. D. Challenges in mapping behaviours to activities using logs from a citizen science project. In: NEXT-GENERATION ANALYST IV, 2016, Baltimore, Maryland. **Proceedings...** [S.l.]: SPIE, 2016. ISBN 9781510600928. 92

MORAIS, A. M. M.; QUILES, M. G.; SANTOS, R. D. Icon and geometric data visualization with a self-organizing map grid. In: COMPUTATIONAL SCIENCE AND ITS APPLICATIONS, 2014, Guimarães, Portugal. **Proceedings...** Switzerland: Springer International Publishing, 2014. p. 562–575. ISBN 978-3-319-09152-5. 77, 78, 79, 91

MORAIS, A. M. M.; SANTOS, R. D. Visualization of citizen science volunteers' behaviors with data from usage logs. **Computing in Science & Engineering**, AIP Publishing, v. 17, n. 4, p. 42–50, 2015. xvii, 23, 25, 26, 41, 74, 77, 79, 82, 83, 84, 85, 86, 91

NEWMAN, G.; WIGGINS, A.; CRALL, A.; GRAHAM, E.; NEWMAN, S.; CROWSTON, K. The future of citizen science: emerging technologies and shifting paradigms. **Frontiers in Ecology and the Environment**, Eco Soc America, v. 10, n. 6, p. 298–304, 2012. 8

NOV, O.; ARAZY, O.; ANDERSON, D. Technology-mediated citizen science participation: A motivational model. In: INTERNATIONAL CONFERENCE ON WEBLOGS AND SOCIAL MEDIA, 5., 2011, Barcelona, Spain. **Proceedings...** Menlo Park, California: The AAAI Press, 2011. 1, 6, 8

O'BRIEN, H. L.; TOMS, E. G. What is user engagement? a conceptual framework for defining user engagement with technology. **Journal of the American Society for Information Science and Technology**, Wiley Online Library, v. 59, n. 6, p. 938–955, 2008. 12, 15, 16, 18, 30, 89, 92

O'NEIL, C.; SCHUTT, R. **Doing data science**: straight talk from the frontline. [S.l.]: " O'Reilly Media, Inc.", 2013. 13, 14, 21

PONCIANO, L. **Computação por humanos na perspectiva do engajamento e credibilidade de seres humanos e da replicação de tarefas**. Tese (Doutorado) — PhD thesis, Universidade Federal de Campina Grande, Campina Grande (2015-11-23 2015), 2015. 53, 54

PONCIANO, L.; BRASILEIRO, F. Finding volunteers' engagement profiles in human computation for citizen science projects. **arXiv preprint arXiv:1501.02134**, 2015. xix, 2, 9, 10, 11, 53, 54

PONCIANO, L.; BRASILEIRO, F.; SIMPSON, R.; SMITH, A. Volunteers engagement in human computation astronomy projects. **Computing in Science and Engineering**, IEEE Computer Society, v. 99, n. 1, 2014. 2, 9, 10, 19, 53

RADDICK, M. J.; BRACEY, G.; CARNEY, K.; GYUK, G.; BORNE, K.; WALLIN, J.; JACOBY, S.; PLANETARIUM, A. Citizen science: status and research directions for the coming decade. **AGB Stars and Related Phenomenastro 2010: The Astronomy and Astrophysics Decadal Survey**, p. 46, 2009. 1, 5

RADDICK, M. J.; BRACEY, G.; GAY, P. L.; LINTOTT, C. J.; MURRAY, P.; SCHAWINSKI, K.; SZALAY, A. S.; VANDENBERG, J. Galaxy zoo: Exploring the motivations of citizen science volunteers. **Astronomy Education Review**, American Astronomical Society, v. 9, n. 1, p. 010103, 2010. 1, 6

RADDICK, M. J.; BRACEY, G.; GAY, P. L.; LINTOTT, C. J.; CARDAMONE, C.; MURRAY, P.; SCHAWINSKI, K.; SZALAY, A. S.; VANDENBERG, J. Galaxy zoo: Motivations of citizen scientists. **Astronomy Education Review**, American Astronomical Society, v. 12, n. 1, p. 010106, 2013. 1, 8

REED, J.; RADDICK, M. J.; LARDNER, A.; CARNEY, K. An exploratory factor analysis of motivations for participating in zooniverse, a collection of virtual citizen science projects. In: INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 46., 2013, Wailea, Maui, HI. **Proceedings...** [S.l.]: IEEE, 2013. p. 610 – 619. ISBN 978-1-4673-5933-7. 8

RIESCH, H.; POTTER, C. Citizen science as seen by scientists: Methodological, epistemological and ethical dimensions. **Public Understanding of Science**, SAGE Publications, p. 0963662513497324, 2013. 7

ROTMAN, D.; PREECE, J.; HAMMOCK, J.; PROCITA, K.; HANSEN, D.; PARR, C.; LEWIS, D.; JACOBS, D. Dynamic changes in motivation in collaborative citizen-science projects. In: CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK, 2012, Seattle, Washington, USA. **Proceedings...** New York, NY, USA: ACM, 2012. p. 217–226. ISBN 978-1-4503-1086-4. 1

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. **Journal of computational and applied mathematics**, Elsevier, v. 20, p. 53–65, 1987. 56

SAAD, M. F.; ALIMI, A. M. Validity index and number of clusters. **International Journal of Computer Science Issues (IJCSI)**, Citeseer, 2012. 57

SHEIKHOLESLAMI, G.; CHATTERJEE, S.; ZHANG, A. Wavecluster a multi-resolution clustering approach for very large spatial databases. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 24., 1998, New York, NY, USA. **Proceedings...** San Fransisco, CA, USA: Morgan Kaufmann Publishers, 1998. p. 428–439. ISBN 1-55860-566-5. 52

SIEBECK, U.; MARSHALL, N.; KLÜTER, A.; HOEGH-GULDBERG, O. Monitoring coral bleaching using a colour reference card. **Coral Reefs**, Springer, v. 25, n. 3, p. 453–460, 2006. 7

SOARES, M. **Employing citizen science to label polygons of segmented images**. Tese (Doutorado) — PhD thesis, Instituto Nacional de Pesquisas Espaciais, São José dos Campos (2011-06-06 2011), 2011. 6, 7

STRUYF, A.; HUBERT, M.; ROUSSEEUW, P. Clustering in an object-oriented environment. **Journal of Statistical Software**, Citeseer, v. 1, n. 4, p. 1–30, 1997. xix, 56

TAN, P.-N.; STEINBACH, M.; KUMAR, V. et al. **Introduction to data mining**. [S.l.]: Pearson Addison Wesley Boston, 2006. 48, 49, 50, 51, 67, 69

VASCONCELOS, L. G. de; BALDOCHI JR., L. A. Towards an automatic evaluation of web applications. In: SYMPOSIUM ON APPLIED COMPUTING, 2012, Trento, Italy. **Proceedings...** New York, NY, USA: ACM, 2012. p. 709–716. ISBN 978-1-4503-0857-1. 92

VESANTO, J. Som-based data visualization methods. **Intelligent data analysis**, Elsevier, v. 3, n. 2, p. 111–126, 1999. 77

VESANTO, J.; ALHONIEMI, E. Clustering of the self-organizing map. **Neural Networks, IEEE Transactions on**, Ieee, v. 11, n. 3, p. 586–600, 2000. 72

VESANTO, J. et al. **Data exploration process based on the self-organizing map**. Tese (Doutorado) — PhD thesis, Helsinki University of Technology, Espoo (2002-05-06 2002), 2002. 70, 72

VESANTO, J.; SULKAVA, M. Distance matrix based clustering of the self-organizing map. In: ARTIFICIAL NEURAL NETWORKS, 2002, Madrid, Spain. **Proceedings...** Berlin Heidelberg: Springer, 2002. p. 951–956. ISBN 978-3-540-44074-1. 72

WANG, W.; YANG, J.; MUNTZ, R. Sting: A statistical information grid approach to spatial data mining. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 23., 1997, Athens, Greece. **Proceedings...** San Fransisco, CA, USA: Morgan Kaufmann Publishers, 1997. p. 186–195. ISBN 1-55860-470-7. 51

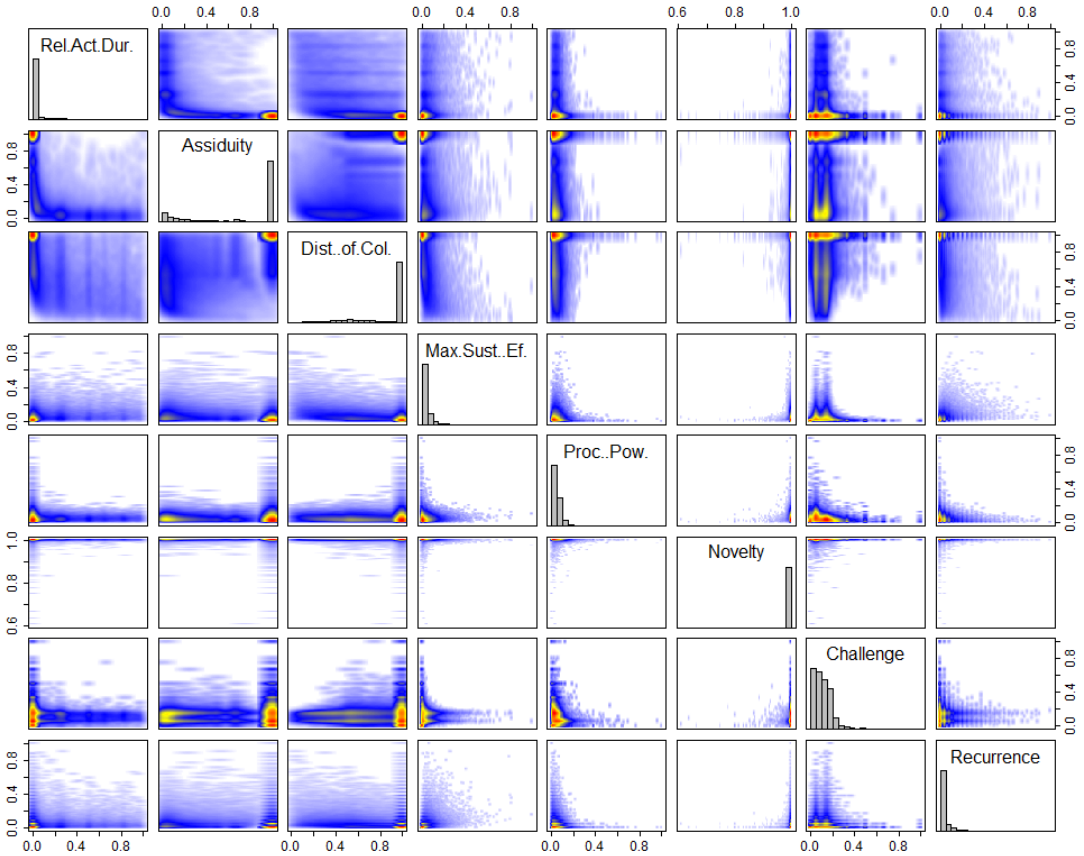
WIGGINS, A.; NEWMAN, G.; STEVENSON, R. D.; CROWSTON, K. Mechanisms for data quality and validation in citizen science. In: INTERNATIONAL CONFERENCE ON E-SCIENCE, 7., 2011, Stockholm. **Proceedings...** [S.l.]: IEEE, 2011. p. 14–19. ISBN 978-1-4673-0026-1. 7

YEUNG, K. Y.; RUZZO, W. L. **An empirical study on principal component analysis for clustering gene expression data**. [S.l.: s.n.], 2000. 37

ADDITIONAL PLOTS

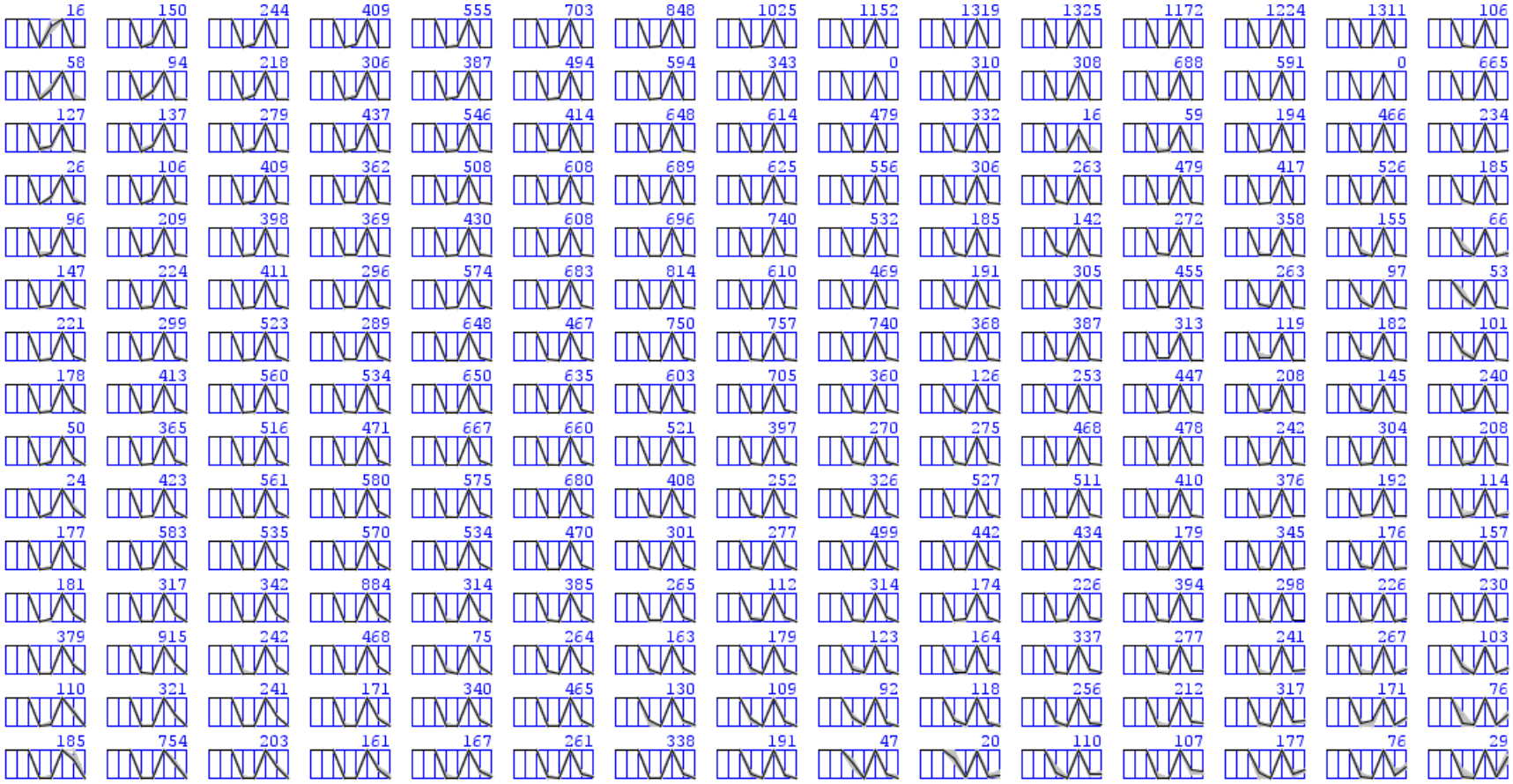
A.1 Looking at Data - Exploratory Data Analysis

Figure A.1 - Scatterplot matrix with estimate data density for whole dataset.



A.2 Data Visualization with SOM grid

Figure A.2 - Visualization of volunteers who collaborate just for one day described by the eight features and standardized by equation 4.4.



PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.