A clustering algorithm to evaluate the attitude of Brazilian researchers regarding open access research data

Bruna S. Freitas[1], Diego Bottero[1], Giancarlo Lucca[1,2], Eduardo N. Borges[1], Helida Santos[1], Graçaliz P. Dimuro[1,3]

[1]Centro de Ciências Computacionais, Universidade Federal do Rio Grande – FURG

[2]Programa de Pós-Graduação em Modelagem Computacional
Centro de Ciências Computacionais, Universidade Federal do Rio Grande – FURG

[3]Departamento de Estadística, Informática y Matemáticas
Universidad Pública de Navarra – UPNA

{bruna.freitas,diegobottero}@furg.br

{giancarlo.lucca,eduardoborges,helida,gracalizdimuro}@furg.br

Abstract. *The core point of the research process are data. They are records from scientific investigation, which support the results published in journals and conferences. Making research data available in open access digital repositories has many advantages, such as increasing the visibility of associated publications, reproducing experiments, and validating results. In Brazil, full and unrestricted sharing of them is not yet accepted by most researchers. This paper presents an initial study to describe a model analyzing the attitude of Brazilian researchers concerning open access research data. A clustering algorithm was used to identify different research profiles. The achieved results indicate the main reasons why the researchers object to share their data.*

1. Introduction

The working group called Brazilian Research Data Network (GT-RDP Brazil) integrates scientists from different research institutes in Brazilian universities. It was formed by the National Education and Research Network (RNP) and the Brazilian Institute of Information in Science and Technology (IBICT). The group considers that the visibility and reputation of Brazilian scientific research are related to the sharing of data collected, generated and used by researchers.

Among the objectives of GT-RDP Brazil, we highlight the following: to identify the practices of Open Access to Research Data in Brazilian institutions and to map the potential national users of the open research data services. In order to achieve this goal, a survey was conducted with researchers from all knowledge areas [Vanz et al. 2018]. A questionnaire was sent to 81,472 Brazilian researchers, including leaders and vice-leaders of research groups registered in the directory of research groups of the National Council for Scientific and Technological Development (CNPq), Graduate Program heads registered in the Coordination for the Improvement of Higher Education Personnel (CAPES), leaders of National Institutes of Science and Technology, and other Brazilian research institutions. A collection containing the questionnaire and the dataset with the 4,703 responses is publicly available in [RDP Brasil 2019].

In [Caregnato et al. 2019], the authors show that, despite the great interest in the topic, there are misconceptions about what sharing and reusing of research data means. It was observed that 49.36% had never used data shared by other researchers and almost a quarter (23.49%) had never shared research data. The sharing of only part of the data produced is a practice pointed out by 53.79% of the researchers. It was also reported that 58.41% do not have an institutional repository available for sharing research data.

The study suggests that the idea of total and unrestricted sharing is not yet accepted by the respondents. However, part of Brazilian researchers indicates receptivity towards data sharing and data reuse. In this context, the main objective presented in this paper is to perform a new descriptive and exploratory analysis of the cited dataset in order to discover different profiles of researchers. Using the ideas given in [Jain 2010] on clustering mining task and with the aid of graphical information visualization, we highlight the main characteristics, behaviors and perceptions of each profile in relation to Open Access to Research Data. The obtained results using K-means++ algorithm [Arthur and Vassilvitskii 2007] provide some understanding on researchers' objections about sharing research data.

The remainder of the paper is organized as follows. Section 2 presents the theoretical background on the algorithms used. In Section 3, we provide the analyzed dataset and each phase of the adopted methodology. Section 4 summarizes the obtained results from the experiments performed. Finally, Section 5 addresses the conclusions and future directions of the presented study.

2. Theoretical framework

In clustering tasks, unlabeled instances or objects are organized into groups based on the similarity of their features. This task goal is to discover the groupings of natural sets of patterns or objects [Jain 2010]. While human beings are excellent at detecting patterns by visual inspection of two- or three-dimensional data, in contexts with high dimensionality, algorithms that automate data clustering are needed. These clustering algorithms maximize intra-group and minimize inter-group similarity. Clustering algorithms can use partitional, hierarchical, or incremental methods.

2.1. K-means clustering algorithm

K-means is a partitional clustering algorithm which divides the dataset into a predetermined number of groups or distinct clusters [Hartigan and Wong 1979, Lloyd 1982]. K-means is formally described in Algorithm 1. Given a set of data objects D and k number of clusters, the algorithm starts to select randomly k objects as initial centroids (line 1). Each object is assigned to the closest centroid c forming k distinct clusters C (lines 2–3). The centroid of each cluster is updated as the average point of objects associated to it (line 4). The process is repeated until the centroids stop changing (line 5). Figure 1 shows an example of the K-means algorithm applied on a two-dimensional dataset, using $k = 3$. In Figure 1 (a) the initial centroids are selected, and in (b), (c), and (d) we see the state of the clusters represented by colors at each iteration.

The main drawback of K-means algorithm is the sensitivity regarding the initiation of centroids. In order to overcome this disadvantage, it was proposed an extension of the algorithm called K-means++ [Arthur and Vassilvitskii 2007]. This extension ensures smarter initialization of centroids and improves the quality of the generated clusters.

Algorithm 1 *K-means*

Input: D set of data objects, k quantity of clusters
Output: C set of clusters
Begin:
1 Select k random objects as initial centroids
2 **Repeat:**
3 Form k clusters C assigning each object to the closest centroid
4 Recalculate the centroid of each cluster C
5 **until** no centroid is changed
End

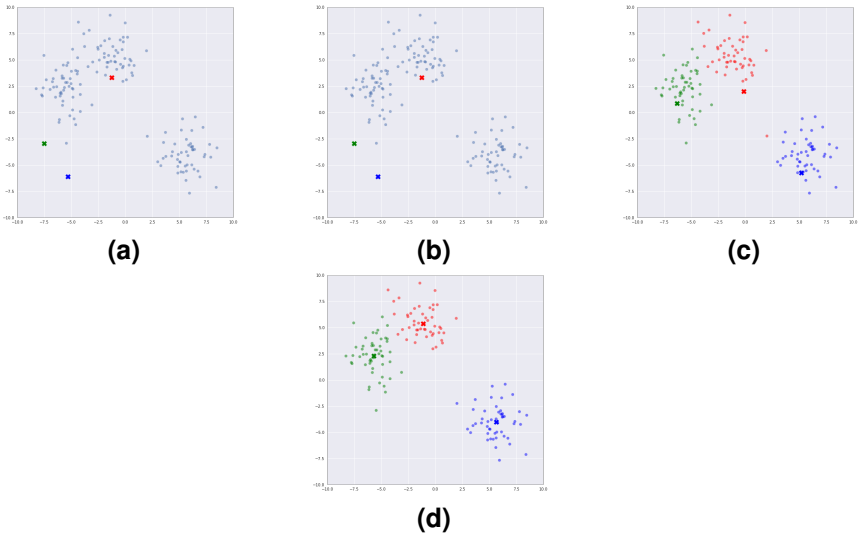

Figure 1. An example of K-means algorithm finding three groups in a two-dimensional data collection.

2.2. Clustering evaluation based on Silhouette coefficient

The optimal number of clusters can be determined using the Silhouette coefficient method [Tomasini. et al. 2017]. For an individual object i, the Silhouette coefficient, represented by $s(i)$, can be calculated as follows:

1. Calculate the average distance $a(i)$ of object i to the other objects in its cluster.
2. Calculate the average distance from the object i to all objects in the other clusters. Find the smallest distance $b(i)$.
3. The coefficient is given by the ratio between the difference of the distances and the maximum distance, according to Equation (1):

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}. \tag{1}$$

The coefficient value can range from -1 to 1. Ideally, the coefficient s should be positive, i.e. $a(i) < b(i)$, and $a(i)$ should be as close to zero as possible, demonstrating high dissimilarity between the groups. Thus, one can find the coefficient for a cluster j with n_j objects by calculating the objects' coefficient average that compose them, Eq. (2).

The average of the clusters coefficients is also applied to the final result of a partitioning into k clusters, according to Eq. (3).

$$s(j) = \frac{1}{n_j} \sum_{i=1}^{n_j} s(i) \tag{2}$$

$$S = \frac{1}{k} \sum_{j=1}^{k} s(j) \tag{3}$$

3. Methodology

The methodology adopted in this study was based on the phases of the Knowledge Discovery in Databases (KDD) process, as illustrated in Figure 2.

3.1. Data selection

The dataset analyzed, entitled "Practices and perceptions on open access to research data", was collected and made available by GT-RDP Brazil, coordinated by RNP and IBICT. The data consist on the responses of an online questionnaire composed of eight sociodemographic questions, thirteen questions about researchers' practices on research data, and other six, concerning the perceptions regarding data sharing and data use. Scientists from research institutions, graduate programs and CNPq research groups were contacted by e-mail and a total of 4,703 valid responses was obtained from all over the country [RDP Brasil 2019][1].

Table 1 presents data on gender and education of the survey respondents. It can be seen that the vast majority of them has a doctoral degree (45.20%) and 36.95% is a post-doctoral researcher. Masters scientists make up 14.91% of the group, and undergraduates account for only 2.94%. In turn, the percentage of female and male respondents were 43.22% and 56.78%, respectively.

3.2. Data preprocessing and transformation

After obtaining the database, it was necessary to preprocess and transform the data in order to remove inconsistencies and improve the reliability of the mining result. In this

[1] `https://hdl.handle.net/20.500.12401/4`

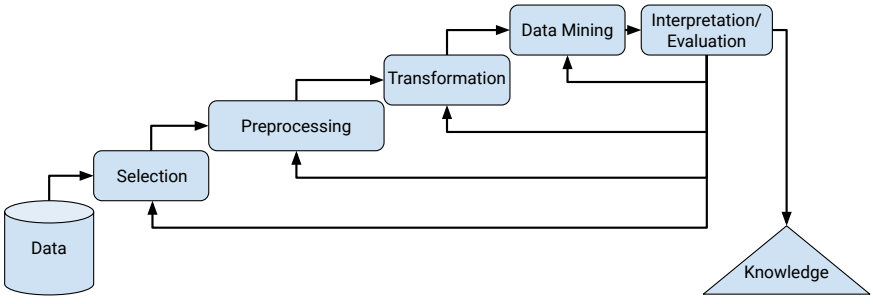

Figure 2. KDD process, adapted from [Fayyad et al. 1996].

Table 1. Education level and gender of the researchers, adapted from [Caregnato et al. 2019].

Education level	Male	%	Female	%	Total	%
Undergraduate	77	60.16	51	39.84	128	2.94
Masters	379	58.40	270	41.60	649	14.91
Doctorate	1092	55.49	876	44.51	1968	45.20
Post-doctorate	924	57.43	685	42.57	1609	36.95
Total	2472	56.78	1882	43.22	4354	100.00

process, Python[2] programming language was used with the aid of Pandas[3] library.

Initially, the dataset consisted of 4694 valid instances and 170 learning attributes. An analysis was performed using the codebook provided by RDP Brazil survey, to understand the codes referring to each learning feature. Thus, the data was obtained through descriptive and objective questions, with single and multiple choices.

All descriptive learning characteristics were removed. Besides, one learning characteristic (Q208) was removed as it contained no description of the answers. Instances associated with researchers who failed to answer at least 15 of the 27 questions were also excluded. Finally, learning features with more than 30% of null values were removed.

Each variable in the preprocessed training set was imputed using the statistical mode, since the variables were categorical. In summary, the transformed data resulted in 3,532 instances with 263 learning features, as the transformation of categorical data to binary data was also applied using the Get Dummies function from Pandas library.

3.3. Data Mining and Validation

The unsupervised clustering algorithm K-means from the scikit-learn[4] library was parameterized with a limit of 300 iterations per run (*max_iter*) and 30 runs with different seeds for the centroids (*n_init*) initialized as in K-Means++ algorithm (*init*). The clustering evaluation method was based on the Silhouette coefficient to validate the formed clusters [Rousseeuw 1987].

In the best achieved clustering, i.e. the partition configuration with the best Silhouette coefficient, some analyses were performed. First, for each relevant question related to different behavioral profiles of researchers, the pattern and distributions of responses in each resulting group were analyzed. These behaviors include: data sharing, use of other researchers' data, data management, concerns about copying and losing publications, misuse, and data use control. Then, chi-square (χ^2) test of independence was also applied [McHugh 2013], in order to statistically substantiate the differences between the groups. This test checks the frequency and relationship between the question responses and the expected distribution, assessing whether or not the sample deviates significantly.

4. Results

Table 2 presents the results of the clustering evaluation. For each value k, it is assigned the clustering Silhouette coefficient. The best result $S = 0.0356$ is achieved for $k = 2$.

[2] https://www.python.org
[3] https://pandas.pydata.org
[4] https://scikit-learn.org

The algorithm split 2,118 (59,96%) instances in group A and 1,414 (40,03%) in group B. The behaviors associated to each group is discussed as follows. We stress that none of the questions were mandatory, so the number of respondents by question ranged significantly.

Table 2. Clustering evaluation by Silhouette coefficient.

k	$s(i)$
2	0.0356
3	0.0294
4	0.0251
5	0.0241
6	0.0203

The analysis on data sharing, data reuse and data knowledge indicates that Brazilian research community (53.79% of respondents) tends to share at least some of their research data [Vanz et al. 2018]. However, the sharing is quite distinct among the identified groups. Figure 3 presents the distribution of responses to the statement "I would share some of my research data in an unrestricted open access repository." The histogram shows the relative frequency of responses for each category in each group, i.e., the axis is presented in percentages. Above each column the absolute frequency of responses is shown. Group A (in blue) shows that the respondents would share at least part of their data without restrictions, since 81.35% of the respondents in this group agree with the statement. On the other hand, for group B (in red) only 16.05% agrees with it. The vast majority of respondents in B group partially agrees (902/1414 = 63.79%) with the statement. The proportion of researchers who disagrees with the statement in group B is also much higher. Thus, regarding open access sharing, it is clear that the behavior of group B is more restricted.

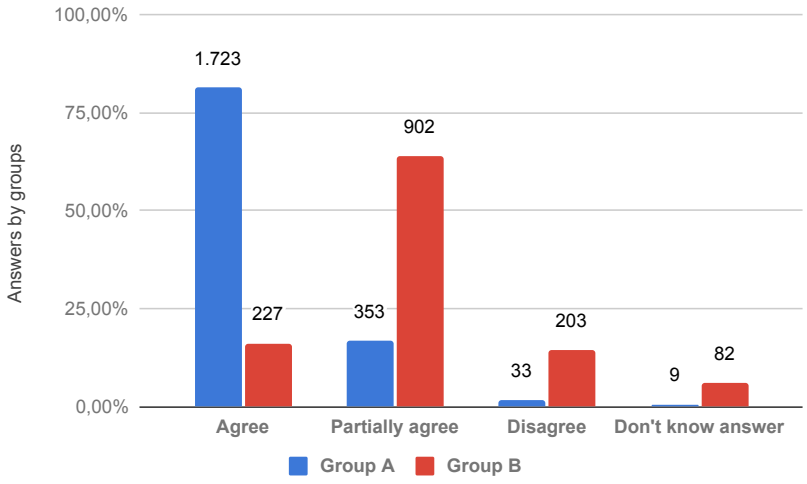

Figure 3. Responses concerning the statement: "I would share some of My research data in an unrestricted open access repository".

Notice that in Figure 3 and in all of the other histograms, the same characteristics regarding data presentation are preserved: categories of answers for the statement/question analyzed, axis and column heights in percentages, absolute frequency above the columns and colors for the groups identified by K-means++.

Regarding data reuse, 49.36% would never use data from other researchers [Caregnato et al. 2019]. The histogram in Fig. 4 shows group A has more experience in using open data (50.8% of respondents). In group B this behavior was only 23.2%.

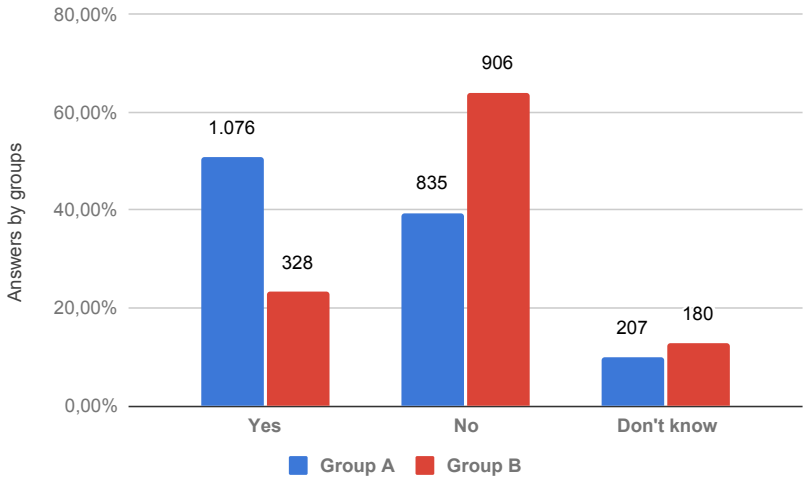

Figure 4. Responses for the question: "Have you ever used open access data shared by other groups in your research?"

Figure 5 presents the distribution of responses for the question "How familiar are you with research data management?". It can be seen that both groups demonstrate knowledge on that, A and B groups with 65.86% and 61.03% of the respondents, respectively. However, it is possible to notice that the knowledge is superficial and balanced between the groups. Therefore, this characteristic did not show significant differences between the researchers of the groups.

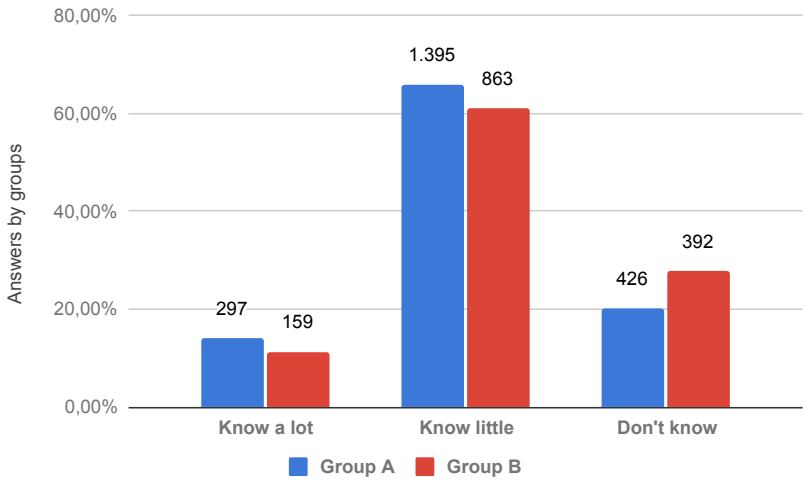

Figure 5. Responses for the question: "How familiar are you with research data management?"

One of the strong obstacles pointed out by the researchers in sharing data is the fear of losing the opportunity to publish the research results. This assumption was presented by 43.89% of the original sample [Vanz et al. 2018]. Analyzing this characteristic across the identified groups, see Fig. 6, the concern is lower in group A, evidenced by the 39.75% of respondents who disagrees with the statement. In group B this proportion drops to less than a half, where only 16.05% disagrees. The relative frequencies of the categories: *I agree* and *I partially agree* are lower for group A, corroborating the pattern more inclined to data sharing.

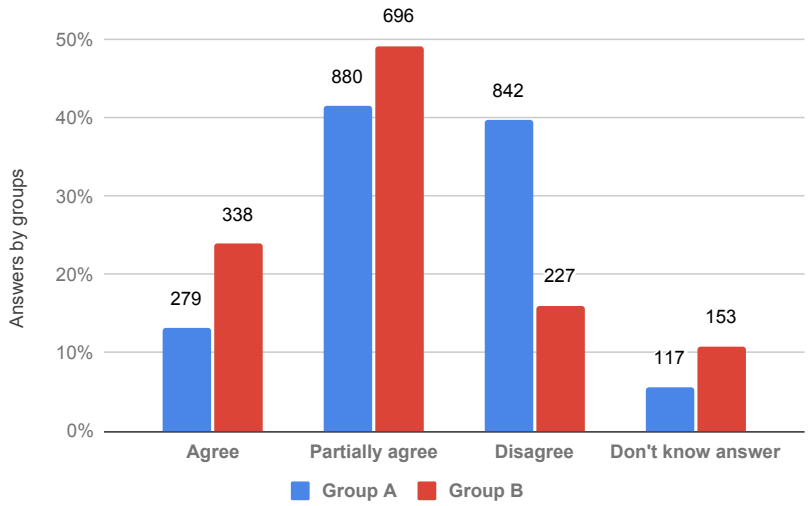

Figure 6. Responses concerning the statement: "I could lose publication opportunities if I shared my data".

Not only the loss of publication opportunities is seen as an obstacle. 47.26% of the researchers also assume that their ideas could be copied if their data were shared. In Figure 7, it can be seen that group B shows greater concern than group A. The relative frequencies for each response category considering groups A / B, respectively, were: 14.68 / 26.66% agree; 46.03 / 51.13% partially agree; 34.94 / 15.56% disagree; and 4.34 / 6.65% did not know how to answer.

Two other troubling issues that stand out in the Brazilian community is the use control of their research data and the assumption that their data would be misinterpreted, probably caused by the dissonance of these thoughts. 45.15% of the respondents always expect for a formal request for the use of their data. The histogram illustrated in Figure 8 shows that the distributions of responses in the categories were similar. The obvious difference is that the proportion who never waits for the request for data use is three times higher in group A than in B. Besides, while 40% in group A always waits for approval, this figure rises to 55.02% in group B.

Furthermore, Figure 9 illustrates group B greater concern about incorrect interpretation and use of their data. 25.04% of the respondents agrees, 54.81% partially agrees, and 12.66% disagrees. In group A, less than a half agrees and more than twice disagrees.

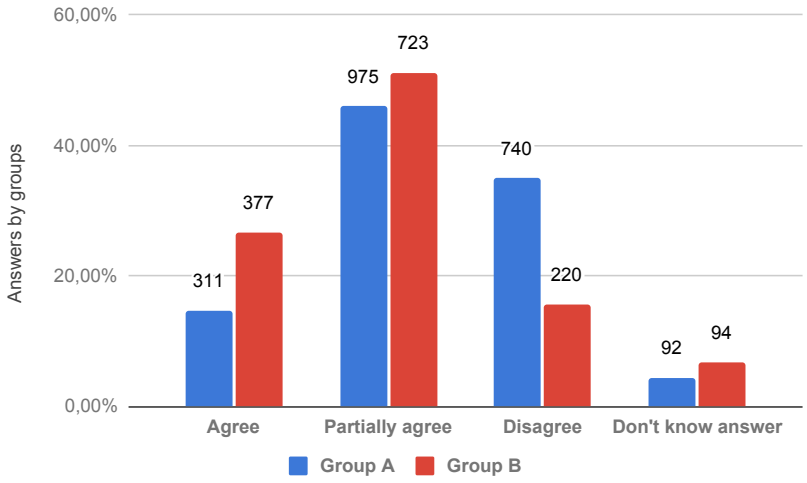

Figure 7. Responses concerning the statement: "My research ideas could be copied if I shared my data".

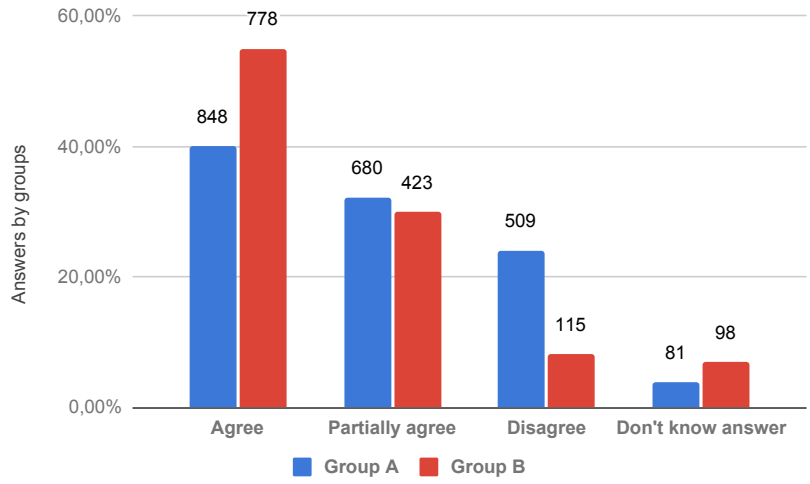

Figure 8. Responses for the statement: "I expect a formal approval request before others use my data".

In order to better understand the discrepancy between groups A and B arose from data mining, chi-square (χ^2) statistical test [McHugh 2013] was performed. This test quantitatively assesses the relationship between question responses and the expected distribution. The values returned by the test are presented in Table 3 and substantiate the differences between the groups, since the p-value returned is less than 0.00001 in all cases, and the statistical coefficient adopted is $\alpha = 1\%$. The questions or statements are ordered with respect to χ^2. The higher the value, the greater the difference between the groups analyzed. For each question the reference of the associated figure is also shown.

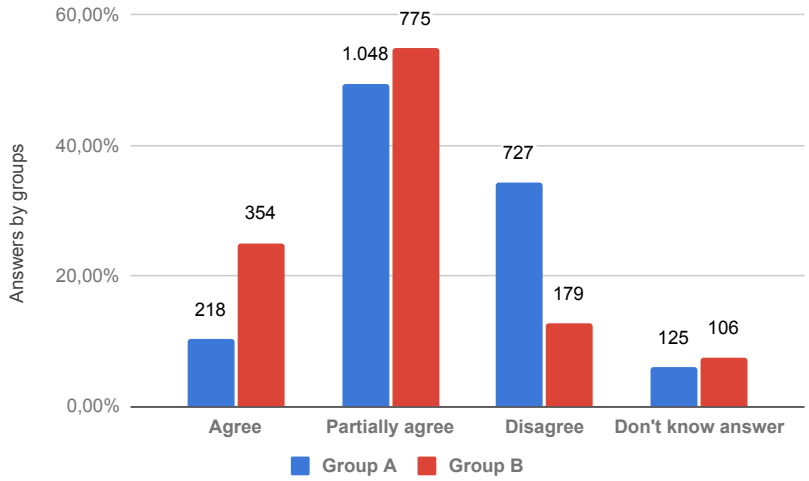

Figure 9. Responses for the statement: "My data could be misused or misinterpreted by other researchers if I shared them".

Table 3. Statistical test result for the analyzed questions, where the p-value is $<$ 0.00001 in all cases.

Fig.	Question / Statement	χ^2
3	I would share some of my research data in an unrestricted open access repository.	1,487.66
9	My data could be misused or misinterpreted by other researchers if I shared them.	276.92
4	Have you ever used open access data shared by other groups in your research?	273.84
6	I could lose publication opportunities if I shared my data.	255.57
7	My research ideas could be copied if I shared my data.	192.76
8	I expect a formal approval request before others use my data.	180.12
5	How familiar are you with research data management?	29.36

Finally, an analysis was conducted regarding the types of research data produced, see Fig. 10. It is important to note that, unlike the other questions, participants were able to tick more than one category response. Researchers from group A generate more observational and documentary data, based on questionnaires and interviews. Group B tends to produce more experimental and observational data. One can conclude that the data characteristics of each group are different. Therefore, this may be a motivating aspect for group B to be less sympathetic to sharing their data, since experimental and observational data often require more generation challenges due to the bureaucracy involved.

5. Conclusions

In this paper, an exploratory analysis of data dealing with behaviors and perceptions of Brazilian researchers about open access research data was presented. In order to conduct

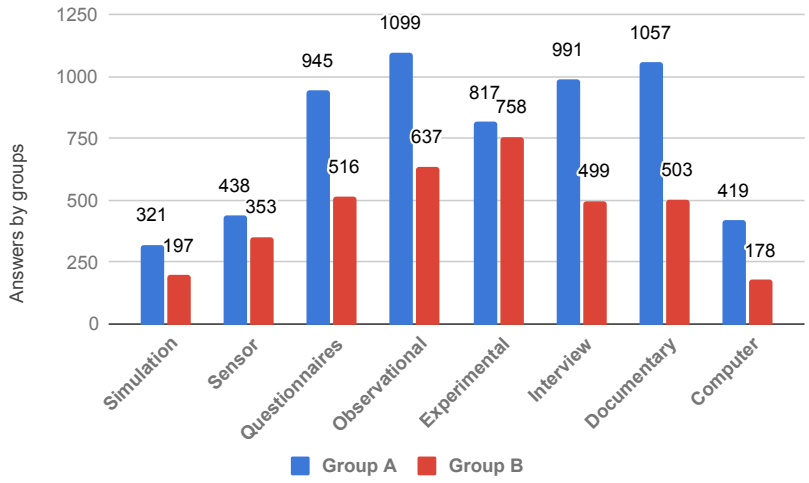

Figure 10. Responses for the question: "Which terms better describe the research data type produced by yourself?".

the study, a real dataset, published by RNP, was used. The adopted methodology, based on KDD, used K-means++ clustering algorithm to train a two-group descriptive model. The obtained results allowed a better understanding of distinct behavioral patterns in the Brazilian research community.

We could observe that group A is predominantly made up of researchers who have already benefited themselves from open access to research data. They claim to have used data shared by others and that they fully agree to share their own data without restrictions. On the other hand, group B is more likely to impose restrictions on their data, such as, demanding formal approval to use them. An interesting fact is that both groups demonstrate some prior knowledge of research data management, even though group B is in minority compared to those who have already shared their data or used others' data. It can be inferred that the restrictions imposed by B group are strongly tied to concerns that their ideas might be copied or that they could miss publication opportunities.

Claiming that, if shared, data would be misinterpreted, may come from lack of knowledge in research data management activities. The data are prepared and documented for the purposes of sharing and improvement, including semantic enrichment, with standardized sets of metadata. Tools for implementing data repositories, such as Dataverse [King 2007] and DSpace [Tansley et al. 2003], are robust and implement all the necessary features for describing datasets and for interoperability of consuming systems.

The results presented in this paper demonstrate the separation of groups with very distinct behaviors, which opens opportunities for further analysis. Future work should include training predictive models for variables of interest, such as those related to data sharing. Besides evaluating the predictive ability, it is essential to understand the models based on interpretable algorithms, like those based on rules and decision trees. In this way, it would be possible to understand which features are more discriminating for certain behaviors and also what the relationships are between these predictive features.

Acknowledgments

This work was supported by the following Brazilian agencies: CNPq (301618/2019-4), FAPERGS (19/2551-0001660, 19/2551-0001279-9) and PNPD/CAPES (464880/2019-00).

References

Arthur, D. and Vassilvitskii, S. (2007). K-means++: The advantages of careful seeding. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, page 1027–1035, USA. Society for Industrial and Applied Mathematics.

Caregnato, S. E., Vanz, S. A. S., Pavao, C. G., Passos, P. C. S. J., Borges, E. N., Gabriel Junior, R. F., Azambuja, L. A. B., and Rocha, R. P. (2019). Práticas e percepções dos pesquisadores brasileiros sobre serviços de acesso aberto a dados de pesquisa. *LIINC em Revista*, 15(2):121–141.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors (1996). *Advances in Knowledge Discovery and Data Mining*. American Association for Artificial Intelligence, USA.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm AS 136: A K-Means clustering algorithm. *Applied Statistics*, 28(1):100–108.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651 – 666.

King, G. (2007). An introduction to the dataverse network as an infrastructure for data sharing. *Sociological Methods & Research*, 36(2):173–199.

Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

RDP Brasil (2019). Práticas e percepções dos pesquisadores brasileiros. Repositórios da Rede Nacional de Ensino e Pesquisa, V2, UNF:6:0pnd8/Eg635y5sVLfSgBrg==.

McHugh, M. (2013). The chi-square test of independence. *Biochemia medica*, 23:143–149.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(1):53–65.

Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., and Smith, M. (2003). The dspace institutional digital repository system: Current functionality. In *ACM/IEEE 2003 Joint Conference on Digital Libraries (JCDL 2003), Houston, Texas, USA, Proceedings*, pages 87–97. IEEE Computer Society.

Tomasini., C., Borges, E. N., Machado, K., and Emmendorfer, L. (2017). A study on the relationship between internal and external validity indices applied to partitioning and density-based clustering algorithms. In *Proc. 19th Int. Conference on Enterprise Information Systems - Volume 3: ICEIS,*, pages 89–98. INSTICC, SciTePress.

Vanz, S. A. S., Passos, P. C. J., Caregnato, S. E., Pavão, C. G., Borges, E. N., Rocha, R. P., Gabriel Junior, R. F., and Azambuja, L. A. B. (2018). Acesso aberto a dados de pesquisa no brasil: práticas e percepções dos pesquisadores: relatório 2018. Available at: http://hdl.handle.net/10183/185195.