Determination of Conformity of Scientific Reports to the Conference’s Topics

Pavel A. Kozlov [1], Shahim I. Safin[1], Vladimir O. Tolcheev [1]

[1] *National Research University "Moscow Power Engineering Institute”, Krasnokazarmennaya 17, Moscow, 11250, Russian Federation*

Abstract

This paper examines the conformity of scientific reports to conference sections. Assumptions were formulated and tested about how scientific articles are allocated to conference sections. With the help of text mining methods, clusters were identified on which the reports are grouped. The analysis of the terminological composition of the obtained clusters is carried out. The closeness of the resulting topics (clusters) and the degree of their correspondence with the names of sections and specialization of departments are analyzed. The sample for research is formed from the proceedings of the interdisciplinary conference "Electronics, Electrical Engineering and Energy", which was held at the National Research University in 2020. The total sample size is 88 articles. Pay attention to an important feature of the data – our sample consists of extremely short text documents, which contain only the titles of reports. In this regard, the text size varies from 5 to 20 words (the average size is 9 words). Our research has shown a fairly significant discrepancy between expert assessments (the selection of a section by an expert) and clusters built using the tool of Data and Text Mining. At the same time the following dependence was established - most often the reports were distributed by the place of work (or study) of the speaker, i.e. the affiliation to the department determined the section. at the same time, the following dependence was established - most often the reports were distributed by the place of work (or study) of the speaker, i.e. the affiliation to the department determined the section. In addition, we have identified high interdisciplinarity of reports, many of which could (with good reason) be reported on several sections at once. Note also that analysis of the specialized literature showed a good correspondence of our results with other similar studies.

Keywords [1]

Data and text mining, clustering, K-means, latent semantic analysis, hierarchical cluster analysis.

1. Introduction

Participation in scientific conferences is an important condition for the successful preparation of qualification works, the implementation of grants, and research work. When choosing conferences, specialists are guided primarily by the information that organizers provide in information letters about the main directions (topics, sections), and assess how the section names are related to their professional interests.

After the presentation of the report, the final decision on its compliance with the profile of the conference, as well as assignment to one of the sections, is made by the experts reviewing the work. However, speaking at a section, the speaker often discovers that his topic is quite different from other works. This can be useful for broadening one's horizons and exchanging ideas from various subject areas, but it does not provide an opportunity to discuss issues with leading experts professionally dealing

[1] Proceedings of VI International Scientific and Practical Conference Distance Learning Technologies (DLT–2021), September 20-22, 2021, Yalta, Crimea

EMAIL: Kozlov.Pavel.Andreevih@yandex.ru (A. 1); safinsi@mail.ru (A. 2); tolcheevvo@mail.ru (A. 3)
ORCID: 0000-0002-9900-9368 (A. 1); 0000-0002-5214-5119 (A. 2); 0000-0002-2320-6533 (A. 3)

with issues of interest to the author. Moreover, quite often works on related topics are presented in different (simultaneously passing) sections of the conference, which makes it difficult to participate in their discussion.

In this article, using the data mining toolkit (Data and Text Mining), the degree of correspondence between the topics of the reports and the titles of the sections are studied based on their terminological proximity [1]. The study is carried out on the example of the analysis of the proceedings of the conference "Electronics, Electrical Engineering and Energy", which was held at the National Research University (NRU "MPEI") in 2020 [2].

2. Study description

The conference "Electronics, Electrical Engineering, and Energy" is interdisciplinary and covers several subject areas, our attention is focused on the direction "Information Technology". This direction is represented by 8 sections, which are conducted by the departments included in the Institute of Information and Computing Technologies of the NRU "MPEI" (ICTI). In our study, 5 sections are considered, which reflect the main areas of training for students studying at the IHTI [1].

- Section 1. "Mathematical Simulation " (Department of Mathematical and Computer Simulation).
- Section 2. "Applied Mathematics" (Department of Applied Mathematics and Artificial Intelligence).
- Section 3. "Computer Science and CAD" (Department of Computational Technologies).
- Section 4. "Computing machines, networks and systems" (Department of Computing machines, systems and networks).
- Section 5. "Management and Informatics in Technical Systems" (Department of Management and Intelligent Technologies).

The total sample size is 88 articles, the number of reports in sections varies from 7 to 23, while each section has on average 2 "external" presentations (except for the section "Mathematical Modeling", which had 8 "external" works). Thus, the topics of the reports for the most part reflect the directions of the department's research.

Let's consider the specifics of presenting materials at the conference "Electronics, Electrical Engineering, and Energy". The authors prepare short abstracts in the following form: "title - text - literature references". The specificity of the original text documents allows the researcher to choose two options: use the entire text (i.e. process full-text data) or analyze only the title of the report (i.e. one sentence). In this paper, it was decided to analyze the titles of reports, the size of which varies from 5 to 20 words (average size is 9 words). Of course, with this approach, the loss of some of the content information located in the main text is possible. However, it is known [2] that the titles of scientific papers well reflect the thematic focus and the main idea of the authors, and the quality of clustering-classification by headings of reports is in good agreement with the results obtained in the analysis of full-text documents. Moreover, recently the processing of individual sentences (and phrases) has become an important area of work in the field of Data Mining, due to the high popularity of various messengers and question-answer systems working with laconic messages.

In this work, a vector model is used for the mathematical approximation of text documents [3,4]:

$$X_j = \begin{bmatrix} x_{1j} \\ x_{ij} \\ x_{Nj} \end{bmatrix} \quad (1)$$

Here N – the number of terms after removing stop words, lemmatization and cutting off single-frequency words, x_{ij} – the weight of term i in document j ($j = 1,..,M -$ the number of documents in the sample, $i = 1,..,N$).

The matrix model was used to describe the sample X:

$$X = \begin{pmatrix} x_{11} & \cdot\cdot & x_{1M} \\ \cdot\cdot & x_{ij} & \cdot\cdot \\ x_{N1} & \cdot\cdot & x_{NM} \end{pmatrix} \quad (2)$$

To date, many different methods have been developed for calculating the weights of terms (weighting by word frequency, tf-idf - weighting, tfc - weighting, etc.). In this work, tfc-weighing will be used, which is a normalized version of the popular and widely used tf-idf-weighing in Data Mining [2,5]:

$$x_{ij} = \frac{f_{ij} * \log(\frac{M}{M_i})}{\sqrt{\sum_{i=1}^{N}[f_{ij} * \log\left(\frac{M}{M_i}\right)]^2}} \qquad (3)$$

Here f_{ij}- frequency of occurrence of the word i in document j and M_i – the number of documents that contain the word i. Summation in the denominator of the fraction is carried out in overall terms of the document j in which the word i occurs.

Our study tests the following assumptions:

- the topics of the reports correspond to the title of the conference section,
- the topics of the reports correspond to the specialization of the department,
- the topics of the reports correspond to the subject area to which several sections belong at once (there is a significant overlap between the topics of the sections and, as a result, similar reports appear in different sections).

Further, in the work, the clustering and visualization of the titles of the reports are carried out, the number of clusters is determined, the name of the cluster is given based on the most frequent terms, the proximity of the resulting clusters and the degree of their correspondence with the names of sections and departments are analyzed [6,7]. Vector models and tfc-weighting of terms are used to represent text documents [8].

3. Analysis of Initial Data Using Data Mining Tools

First of all, we will visualize the initial sample based on metadata, using an expert assessment of the reports belonging to a section. For visualization, we will apply the method of principal components (PCA) [3,4].

The visualization result, shown in Figure 1, shows the absence of obvious clusters corresponding to the sections and suggests the presence of reports that are "close" to several topics at once, as well as works that are uncharacteristic (atypical) for the sample under consideration.

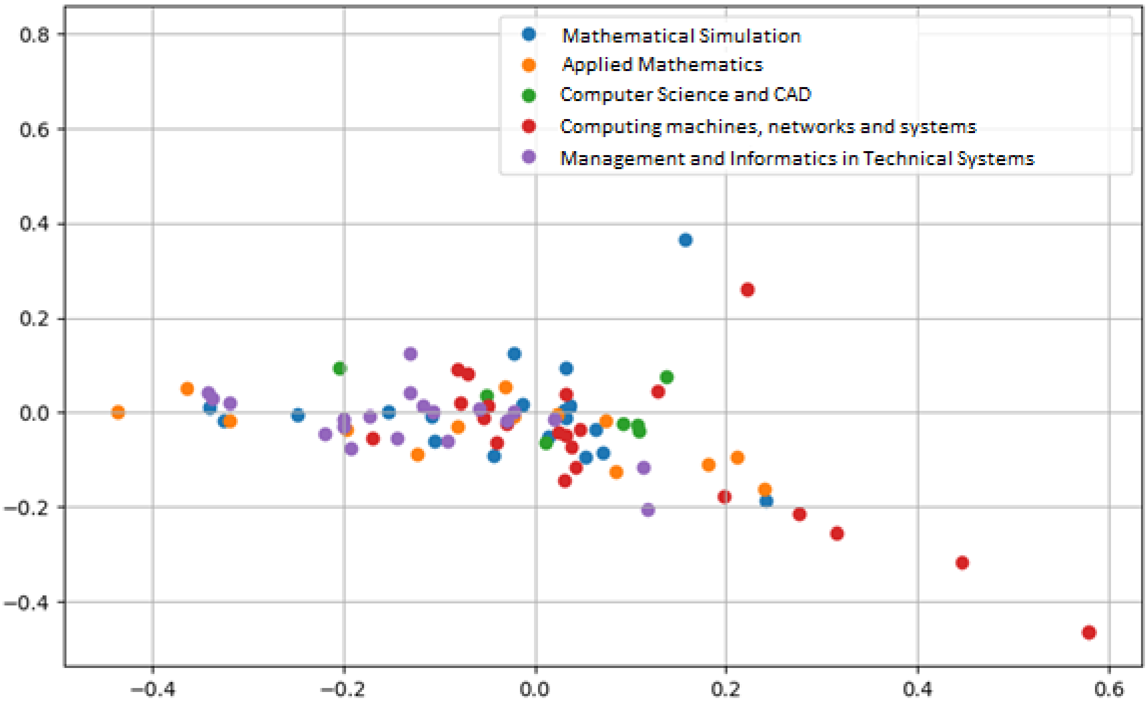

Figure 1: Projection of a sample into two-dimensional space with PCA

Let us formulate the following assumption: each section has a “core” of documents that are quite close to each other (and, in general, correspond to the section name), and a “periphery” when the texts differ from the main topics. To test this assumption, we use hierarchical cluster analysis, which allows us to analyze in more detail the original sample using a dendrogram constructed using the cosine measure of proximity and combining clusters using a weighted pairwise average [9].

Figure 2 shows a dendrogram that was obtained using the cosine measure of proximity and with the union method - weighted pairwise average. The results presented in Figure 2 only partially confirm the earlier assumption. It can be seen that there are fairly close articles that are combined into (very) small clusters (from 2 to 5 articles). In general, the source data can be combined into 5 large clusters. However, the “model” of such clusters will not correspond to our assumption of the presence of a “core” and “periphery”. The result of the hierarchical cluster analysis and the visualization performed suggest that the "model" is a combination of small groups of (related) documents distant from each other into a rather heterogeneous formation (a blurred cloud stretched in the feature space).

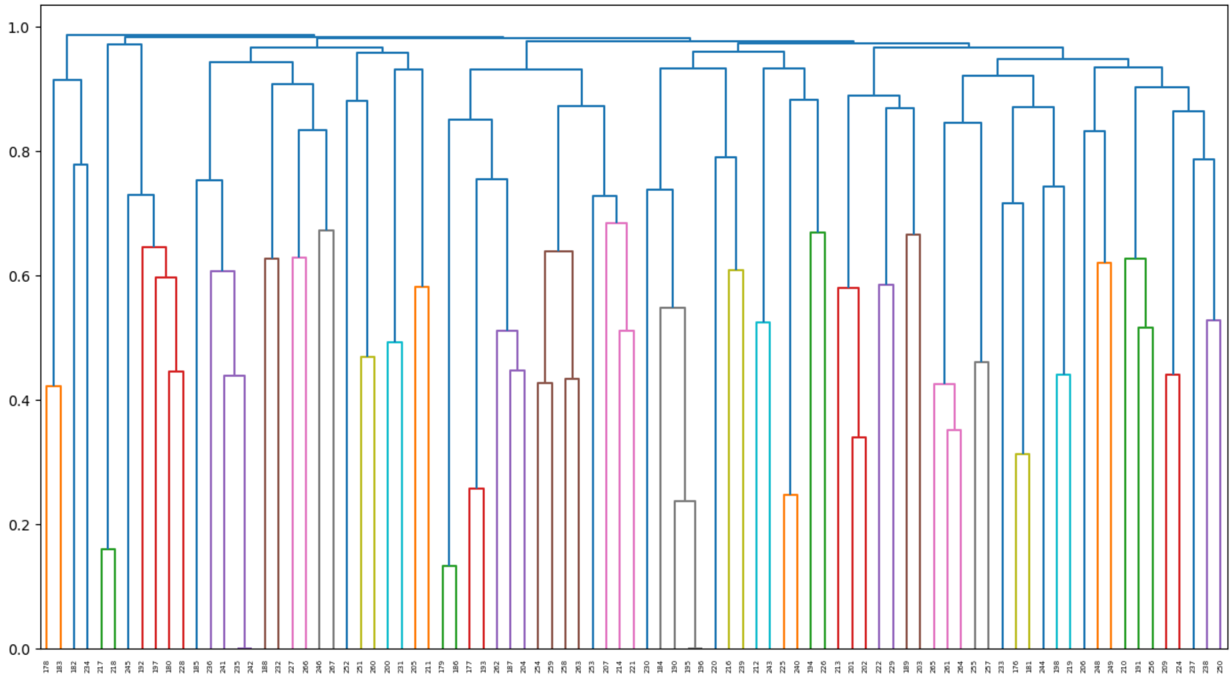

Figure 2: Hierarchical clustering of the original sample

The next study includes the use of data mining methods for the analysis of the initial sample without taking into account expert assessments. An automatic division of a set of reports into terminologically close groups (clustering problem) is considered. To carry out clustering, we use the method of K-means and latent semantic analysis (LSA) [10,11]. As before, we are interested in the distribution of documents into five clusters (the number of sections) using data mining.

The clustering results presented in Figures 3 and 4 allow us to conclude the presence of 5 topics of reports, which corresponds to the initial number of sections.

Let's analyze the grouping of articles and give a name to the resulting clusters:

1. Subject "Simulation". High-frequency cluster words – simulation, modeling, model, method.
2. Subject "Computer systems". High-frequency cluster words - system, method, detection, research, implementation, computational.
3. Subject "Spacecraft control systems". High-frequency words of the cluster - system, control, apparatus, analysis, space.
4. Subject "Algorithms". High-frequency cluster words - algorithm, implementation, research, method, analysis.
5. Subject "Information systems". High-frequency words of the cluster - system, control, decision, information, process.

Only one cluster ("Simulation") coincided with the original section names. The rest of the clusters turned out to be "interdepartmental" and "intersectional". On the whole, they characterize rather well the direction of the ICTI activity, reflecting the high intersectionality of research. Along with well-interpreted constellations, a rather unexpected cluster associated with spacecraft control has also formed. Among the main specializations of the departments of ICTI, there are no aerospace topics. Let's analyze the resulting cluster in more detail. It consists of five articles, which were reported in the sections "Computing machines, networks and systems" and "Control and informatics in technical systems." The work was carried out by different research teams on the following topics: "research of data filtering methods", "motion control using solar sensors". When the number of clusters is reduced to three, reports on "spacecraft" are added to the "Information systems" cluster. To check the "stability" of the aerospace cluster, it is necessary to analyze the proceedings of the next conferences and estimate the number of papers submitted annually.

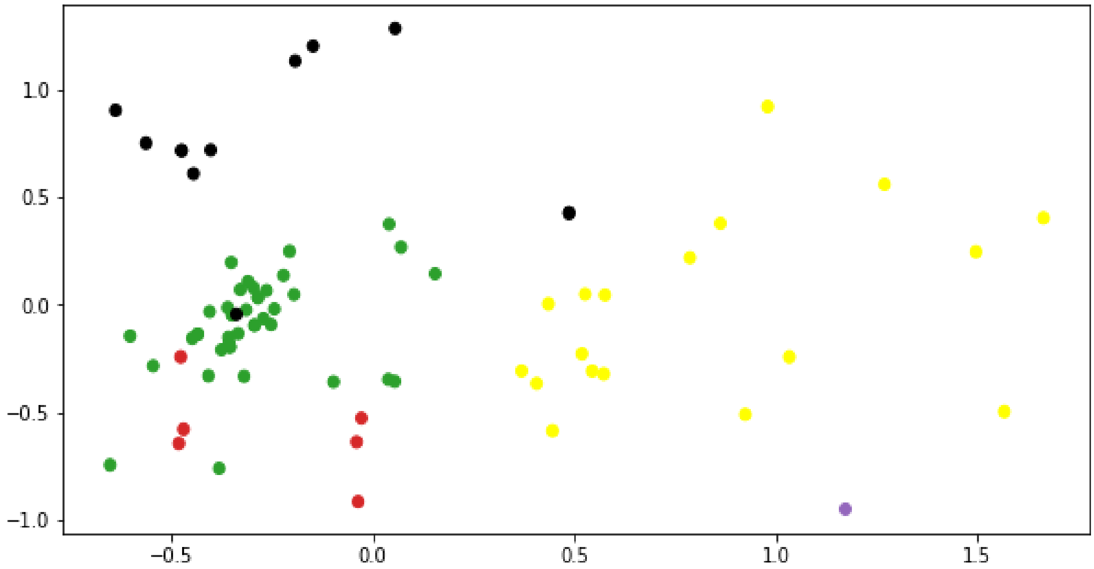

Figure 3: K-Means Clustering

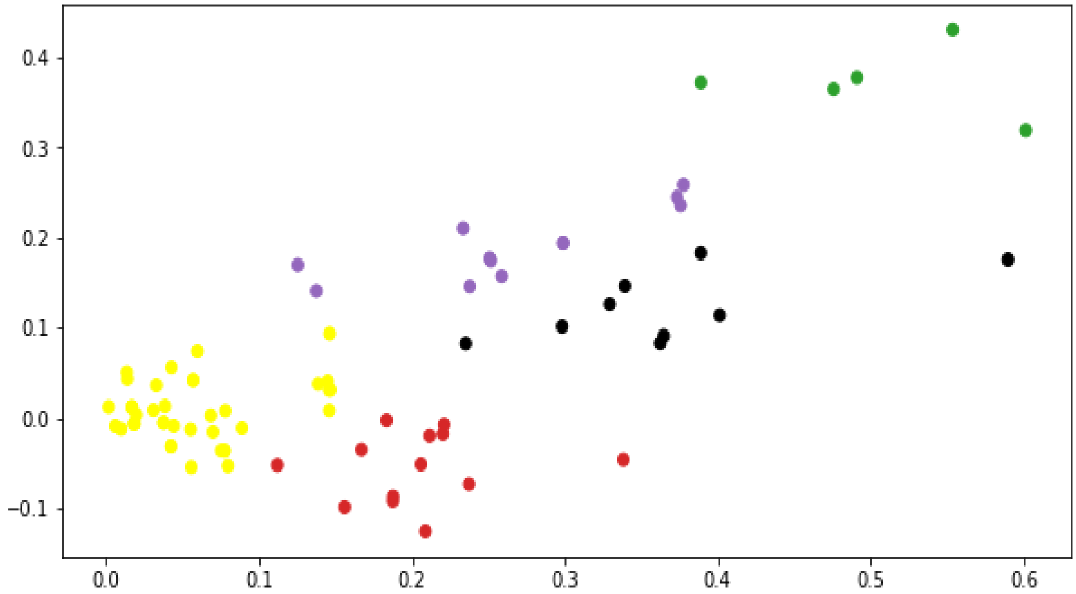

Figure 4: Clustering by LSA method

Additionally, analyzing the initial distribution of articles, the following conclusions can be drawn:

- The topics of reports in half of the cases (53%) correspond to the title of the conference section. Also, here we can talk about the essential interdisciplinarity of the sections, for example, the directions "Applied Mathematics" and "Mathematical Simulation" are close. Some topics can fit most of the sections, for example, "Neural Networks".

- The themes of the reports correspond to the specialization of the department. For the department "Mathematical Simulation" there is a great correspondence between the topics of reports and specialization, for other departments a high level of compliance is also revealed. This suggests that authors are more likely to report at those sections that their department heads, and not at those that are optimal for them in terms of the topic.

4. Conclusions

The studies carried out show that the initial topics of the reports do not always coincide with the topics of the section, but most often correspond to the specialization of the department (in particular, those carried out at the department of research and development). The results obtained are in good agreement with similar works [12,13,14,15]. So, in [13], using LSA, the correspondence of the titles of reports and sections of conferences "Mathematical Methods of Pattern Recognition" was checked based on the analysis of bibliographic descriptions (titles, annotations, keywords). The automatically constructed clusters differed significantly from the original headings of the conference. It seems that the distribution of scientific articles by sections of the conference is most often carried out quite subjectively and is not confirmed by the results that are obtained when using the data mining tools (in particular, when carrying out automatic clustering) [12,13,15]. Of course, the exclusion of expert assessments from the process of distributing articles into sections is hardly advisable. However, the use of combined approaches, including, along with expertise, the use of data mining means, will contribute to the formation of sections "declared" by the organizers and increase the thematic proximity of the works reported at one session.

5. References

[1] Collection of abstracts of the XXVI international scientific and technical conference of students and graduate students Radio electronics, electrical engineering, and energy. https://reepe.mpei.ru/Pages/default.aspx.

[2] G. Salton, The SMART retrieval system: Experiments in automatic document processing. Englewood Cliffs, N.J.: Prentice-Hall, 1971. DOI:10.1109/TPC.1972.6591971.

[3] C.C. Aggarwal, Machine Learning for Text. Springer, 2018. 452 p. DOI:10.1007/978-3-319-73531-3.

[4] Flach P. Machine learning – The Art and Science of Algorithms that Make Sense of Data. Cambridge University Press; 1st edition (September 1, 2012) DOI:10.1017/CBO9780511973000.

[5] K. Chen, Z. Zhang, J. Long, H. Zhang, Turning from TF-IDF to TF-IGM for term weighting in text classification, Expert Syst. Appl. 2016, 66, 245–260.

[6] S.R. Nair, G. Gokul, A.A. Vadakkan, A.G. Pillai, M.G. Thushara, Clustering of research documents – a survey on semantic analysis and keyword extraction. 6th International Conference for Convergence in Technology, Maharashtra, India, 2021. DOI:10.1109/I2CT51068.2021.9418197.

[7] R. Lakshmi, S. Baskar, Efficient text document clustering with new similarity measures. International Journal of Business Intelligence and Data Mining. 2021. V. 18. № 1. PP. 109-126 DOI:10.1504/IJBIDM.2021.111741.

[8] K. Aas, L. Eikvil. Text Categorization: A Survey. Norwegian Computing Center. Oslo. 1999, p.1–37.

[9] P.Kozlov, A. Mokhov, V. Tolcheev, Detection of the thematic groups in scientific publications, Russian Advances in Fuzzy Systems and Soft Computing: Selected Contributions to the 8th International Conference on Fuzzy Systems, Soft Soft Computing and Intelligent Technologies, FSSCIT 2020; Smolensk; Russian Federation; 29 June 2020 до 1 July 2020; Код 165822. Volume 2782, 2020, Pages 278-285

[10] D. Zeebaree, H. Haron, A. Abdulazeez. Combination of K-means clustering with Genetic Algorithm: A review. International Journal of Applied Engineering Research, Volume 12, Number 24, 2017, pp. 14238-14245.

[11] C. Li, Y. Lu, J. Wu, Y. Zhang, Z. Xia, P. Liu, T. Wang, D. Yu, X. Chen, J. Guo. LDA meets Word2Vec: a novel model for academic abstract clustering. Web Conference, 2018, p. 1699-1706.

[12] A.A. Kuzmin, A.A. Aduenko, V.V. Strijov, Thematic Classification for EURO/IFORS Conference Using Expert Model, Conference of the International Federation of Operational Research Societies, 2014, p. 175.

[13] M. V. Korotchenkov, E. Kh. Khunov, Identification of topics of scientific documents based on latent semantic analysis, Radio electronics, electrical engineering, and energy: Abstracts of the twenty-seventh international scientific and technical conference of students and graduate students, Moscow, NRU "MPEI ", 11-12 March 2021, p. 254.

[14] S. Henry, B.T. McInnes. Literature-based discovery: models, methods, and trends. Journal of biomedical informatics, vol.74, 2017, pp.20-32. DOI: 10.1016/j.jbi.2017.08.011.

[15] M. Kamada, M. Isonuma, K. Asatani, I. Sakata, Discovering Interdisciplinarily Spread Knowledge in the Academic Literature. DOI:10.1109/ACCESS.2021.3110111.