fire

MDPI

Article

Text Mining Approach for Trend Tracking in Scientific Research: A Case Study on Forest Fire

Yunus Eroglu

Industrial Engineering Department, Faculty of Engineering and Natural Sciences, İskenderun Technical University, 31200 İskenderun, Hatay, Türkiye; yunus.eroglu@iste.edu.tr

Abstract: Scientific studies are increasing day by day with the development of technology. Today, more than 171 billion academic records are made available to researchers via the Web of Science database, which is frequently followed by the scientific community, and is where records of articles, proceedings, and books in many different fields are kept. More than 40 thousand studies are reached when a search is made for research on forest fires in the relevant database. It is unfeasible to examine and read so many publications and understand what topics are important in the relevant field, what is trending, or whether there is a difference between the subjects studied based on years and/or regions/countries. The most effective and scientific method of deriving information from such large and unstructured data is text mining. In this study, text mining is used to reveal where the research on forest fires in the Web of Science database concentrates, which study topics have emerged, how an issue's level of importance changes over the years, and which topics different countries focus on. Therefore, the abstracts of approximately 32 thousand articles published in English were collected and analyzed based on the country of the authors and the published years. Over 600 words in the abstracts were indexed for each article and their importance was calculated according to inverse document frequency. A size reduction was made to determine the main concepts of the articles by using the singular value decomposition and a total of 29 different concepts were found. Among these, important concepts can be mentioned such as damage to vegetation and species affected, post-fire actions, fire management, and post-fire structural changes. Considering all the articles, studies on soil, fuel (biofuel), treatment, emissions, and species were found to be important. The results we have obtained in this study are by no means a summary of the research carried out in the field; they do, however, allow statistical due diligence concerning, for example, which subjects are important in the relevant field, the determination of increasing and decreasing trending topics, which countries attach importance to in the same research, and so on. Thus, it will function as be a guide in terms of the direction, timing, and budget allocation of research plans in a specific area in the future.

Keywords: topic extraction; trend analysis; text mining; forest fire

Citation: Eroglu, Y. Text Mining Approach for Trend Tracking in Scientific Research: A Case Study on Forest Fire. *Fire* **2023**, *6*, 33. https://doi.org/10.3390/fire6010033

Academic Editor: Grant Williamson

Received: 8 December 2022
Revised: 31 December 2022
Accepted: 9 January 2023
Published: 13 January 2023

1. Introduction

There are approximately 3.5 billion hectares of forested area around the world. A total of 119 million hectares of tree cover have been lost from fires globally in the last two decades. Forest fires are observed in many parts of the world, day by day, and their damage to the ecosystem is increasing. Forest fires occur in every continent and country from Australia to Canada, from the United States to China, damaging the environment, wildlife, human health, and infrastructures of regions [1]. Table 1 provides information on tree cover lost due to fires in various countries from 2001 to 2021 [1]. The countries in this table are selected as the countries that have published one thousand or more articles in the Web of Science database on the topic of forest fires. Between the years 2001–2021, 52.8 million hectares forest area was lost due to fires in Russia. Looking at the total forest area losses, Russia, the USA, Brazil, Australia, and Canada lost over one million hectares.

Australia and Canada lost more than 7% of their total forest assets from fires between the years 2001 and 2021.

Table 1. Stable forests and lost forest area statistics by country between the years 2001–2021 [1].

Country	Stable Forests (Mha)	Burned Forest Area 2021 (kha)	Burned Forest Area 2001–2021 (Mha)	Burned Forest Area 2001–2021 (%)
Australia	80.3	132	6.26	7.796
Brazil	414	596	9.51	2.297
Canada	257	1610	2.68	1.043
England	1.2	0.021	0.007	0.583
France	15.8	3.51	0.048	0.304
Germany	11.7	0.476	0.006	0.051
Italy	9.26	6.3	0.045	0.486
China	202	22.7	0.893	0.442
Russia	686	5360	52.8	7.697
Spain	11.4	16.1	0.300	2.632
USA	238	846	11.1	4.664

Forest fires are so crucial that many institutions and organizations contribute to studies related to this field by providing funds. According to the Web of Science database, which is widely used for academic research, more than 20 thousand studies on Forest Fires (over 40 thousand so far) have been funded by various institutions or agencies [2]. In Table 2, the agencies that have funded studies on this subject and how many studies they have funded so far are given. As expected in relation to forest fire statistics, the majority of funding agencies originate from countries that lost their forests due to fire.

Table 2. Agencies funding studies on forest fire [2].

Funding Agencies	Record Count	%
National Science Foundation NSF	2067	5.05
European Commission EC	1461	3.57
United States Department of Agriculture USDA	1406	3.43
United States Forest Service USFS	1120	2.73
National Natural Science Foundation of China NSFC	1001	2.44
Natural Sciences and Engineering Research Council of Canada NSERC	941	2.30
Spanish Government	784	1.91
Conselho Nacional de Desenvolvimento Cientifico e Tecnologico CNPQ	649	1.58
UK Research Innovation UKRI	640	1.56
National Aeronautics Space Administration NASA	608	1.48
Natural Environment Research Council NERC	506	1.23
Australian Research Council	505	1.23
United States Department of Health Human Services	476	1.16
National Institutes of Health NIH USA	461	1.12
Portuguese Foundation for Science and Technology	436	1.06
Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior Capes	434	1.06
Joint Fire Science Program	406	0.99
CGIAR	372	0.90
United States Department of Energy DOE	354	0.86
Russian Foundation for Basic Research RFBR	333	0.81
Australian Government	332	0.81
German Research Foundation DFG	332	0.81
Ministry of Science and Innovation Spain MICINN	255	0.62
Fundacao de Amparo a Pesquisa do Estado de Sao Paulo FAPESP	251	0.61
NSF Directorate for Biological Sciences BIO	249	0.60

Forest fire is an ecological threat for the entire world. The forest fire statistics reveal the importance of allocating the funds provided by many agencies to the right studies in the relevant field. The crucial topics in any field, the concepts with increasing or decreasing trends, or current and obsolete areas of research can be revealed by examining the research conducted in the that area. Such a study is primarily made possible through a detailed

search of the relevant literature. As seen in Figure 1, studies on forest fires are increasing exponentially from year to year [2]. Given the massive growth in forest fire literature, a systematic approach is required to review these publications. However, it is not possible to systematically review so many publications manually; many problems may arise. The biggest problem is unstructured textual information that multiplies over a short time.

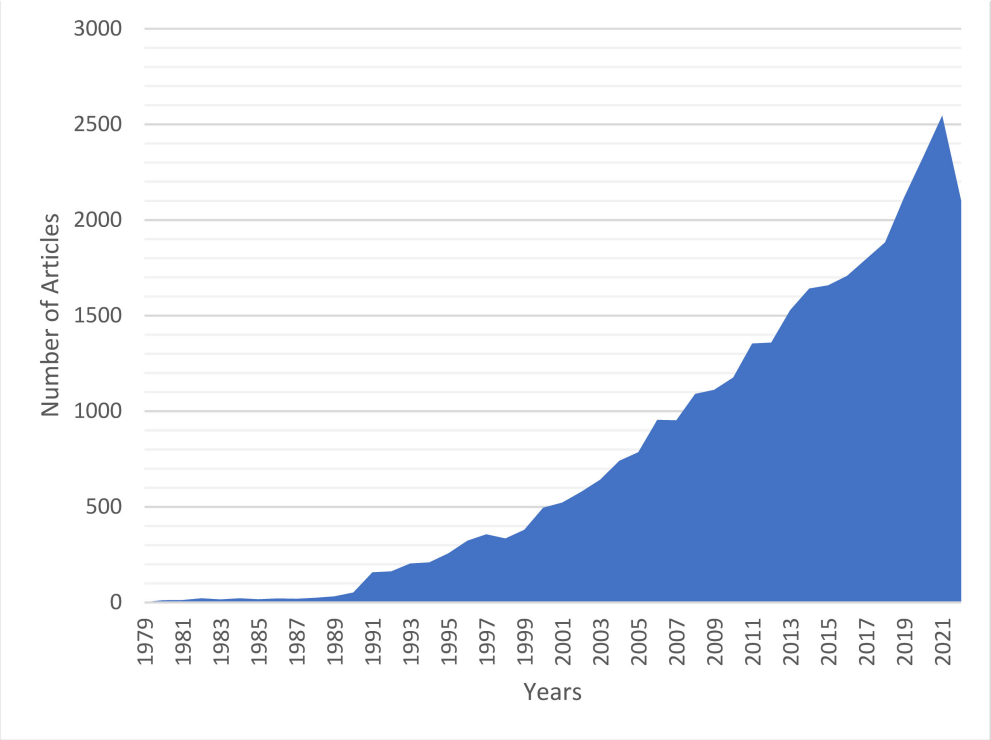

Figure 1. Number of published articles on forest fires by year.

Complex systematic reviews can take more than a year to complete, and half the time can be spent searching and ensuring one is scanning correctly. This poses a major problem; policy-makers and practitioners often want research results on a much shorter time scale than manual methods would allow [3]. In this study, a systematic review with text mining is suggested for the current literature review—a task that is almost impossible to do manually—in order to determine the important topics and trends in the field of forest fire. In this way, it is possible to identify hot topics, frameworks, policies, future challenges, and expectations that are of worldwide importance.

Knowledge discovery from structured data by data mining has recently attracted the attention of a considerable number of researchers and industry professionals [4]. The main reason for the widespread use of data mining techniques is that data can be easily stored and processed in digital media; the manual processing and interpretation of highly derived data is expensive and time consuming. Although the data commonly used in data mining studies are structured and numerical, various research articles, comments, and discussions may contain unstructured textual data about a specified area. Text mining, a method of data exploration, is a trending tool to help eliminate errors, save time, and provide more precise information from unstructured texts. It uses some word indexing methods to exploit the enormous amount of information contained in text documents [5]. Many techniques developed for text mining in the literature are widely used for knowledge discovery in both academia and industry [6].

There are many kinds of industrial studies on retail, pharmaceutical, aviation, and energy for purposes such as social media analysis, customer satisfaction, business intelligence, or identifying crucial topics. Shim et al. (2021) investigated Korean response to COVID-19 vaccines on social media [7]. Their main purposes were investigating the

emotional attitude of the public against vaccines and identifying prominent topics related with COVID-19 vaccines. Thus, they collected the posts containing the words "coronavirus" and "vaccines" on twitter and analyzed them via text mining. Their results emphasized that there was an increase in tweets reporting anger and opposition with the increase in the number of cases. Kitsosis et al. (2021) presented a customer satisfaction study on hotel guests to determine specific product attributes or service characteristics [8]. In another study, Atay et al. (2021) mined the current literature in the field of airline optimization to understand trends and commercial threats in the airline industry [9]. Their results showed that more robust and consistent aviation in the future will depend on paying more attention to airline optimization studies. Moreover, they emphasized that new methods and technologies should be developed in airline operations to minimize the damage by considering environmental factors. As for examples from the field of energy, text mining research in the literature for the trend of optimization studies in wind energy [10] and the analysis of wind turbine accident news can be examined as examples [11]. In line with the research of the authors, almost no study was found using the text mining method related to forest fires; the exception was a study based on the analysis of tweets posted after the forest fire [12]. Mustaqim et al. (2020) presented a text mining study of twitter posts to analyze the public's thoughts about the government's practices after a forest fire in Indonesia.

In this study, the aim is to discover important and trending topics by examining academic research focused on "forest fire" using the text mining method. Our research results are expected to guide the determination of new research topics. Absences and prominent issues in the literature have been provided by mining of the enormous and increasing number of increasing articles in a fast, systematic, and subjective way. Consequently, it will be possible to allocate funds correctly by giving priority to more important and trending issues.

2. Materials and Methods

The primary tasks of text mining are to identify the trends in academic studies on forest fires that have been made so far, and to identify the key issues that lost their importance or gained importance. Text mining starts with data gathering. Then, collected data needs preprocessing to digitize unstructured textual data. Thereafter, word indexing is conducted. The resulting word indices now constitute structured and numerical data; hence, they are suitable for data mining. Hereafter, the results obtained are ready to be analyzed and interpreted.

2.1. Data Gathering

In this study, approximately 32 thousand articles (as of 22 November 2022) in the Web of Science database containing "Forest Fires" or "Wildfires" in their title, summary, or keywords were analyzed. The first step applied in obtaining this data is to filter the articles with "forest fires" or "wildfire" in their topics from the whole database. Afterwards, only the articles as the document types were filtered, the others (proceeding papers, letters, and corrections, etc.) were excluded from the research. Then, in order to observe the country-based results, filtering was conducted by country. During this filtering process, it was found that there were articles from 175 different countries/regions. Considering that it would be time-consuming to examine all countries one by one, and considering that there are ten or less articles from some countries, it was decided to conduct the analysis in groups for countries. Countries were analyzed in four groups. Countries with more than a thousand publications were analyzed separately. Table 3 gives the number of publications by countries which have more than 1000 publications. Group A contains countries which have publications between 500 and 999 (see in Table 4), Group B contains countries which have publications between 100 and 499 (see in Table 5), and other countries are grouped as Group C for fewer than 100 publications. In group C, there are 134 countries and 45 of them have fewer than 10 publications.

Table 3. Number of publications by countries which have more than 1000 publications.

Country	Number of Publications
USA	14,181
Canada	3732
Australia	3214
Spain	2567
Germany	1828
Peoples R. China	1802
England	1703
Brazil	1521
France	1436
Italy	1155
Russia	1011

Table 4. Number of publications by countries which have publications between 500 and 999 (Group A countries).

Country	Number of Publications
Portugal	998
Sweden	823
India	727
Netherlands	724
Japan	688
Switzerland	655
Finland	553
Greece	536
South Africa	508

Table 5. Number of publications by countries which have publications between 100 and 499 (Group B countries).

Country	Number of Publications	Country	Number of Publications
Indonesia	464	Poland	248
Argentina	406	Israel	229
Mexico	394	Iran	197
Chile	329	Czech Republic	190
South Korea	324	Wales	185
Scotland	323	Denmark	157
Norway	304	Malaysia	151
New Zeland	297	Singapore	151
Australia	288	Thailand	147
Belgium	271	Taiwan	100
Turkey	267		

Thus, abstracts of the articles were obtained separately for eleven different countries and separately for three different groups of countries. Although around thirty-two thousand articles were filtered when only the subject was searched, more than forty-eight thousand article abstracts were obtained as a result of country-based filtering. The main reason for this situation is that many articles were written by multinational research teams. This means that the same article is obtained repeatedly in different country filters. In the data obtained, there is a unique Web of Science (WOS) number for each article. The articles were checked according to their WOS numbers and the repetitive articles for different country groups were labeled as Multinational in the data obtained. Thus, for example, if a USA-based study was also obtained in a different country filter (a researcher in the USA and a researcher in a different country published jointly), that study is now included

in the multinational category. This situation has caused differences (i.e., decreases) in the number of publications of the countries in the last database created and arranged for analysis. Figure 2 shows the final distribution of studies by countries that are the addresses of the authors. The USA takes the first place with more than nine thousand papers when the author addresses of the studies on forest fires are examined. Because of the increase in online working opportunities, the studies carried out by scientists working in different countries (multinational) are close to ten thousand. In addition, when the data in Table 3 and Figure 2 of the USA-based studies are compared, it is understood that almost 30% of them are conducted with multinational co-authors. The main reason for the difference in the number of country-based publications in Table 3 and Figure 2 is the fact that the studies conducted with co-authors and grouped as multinational. The sum of multinational and the USA-addressed publications constitutes approximately 60% of all publications.

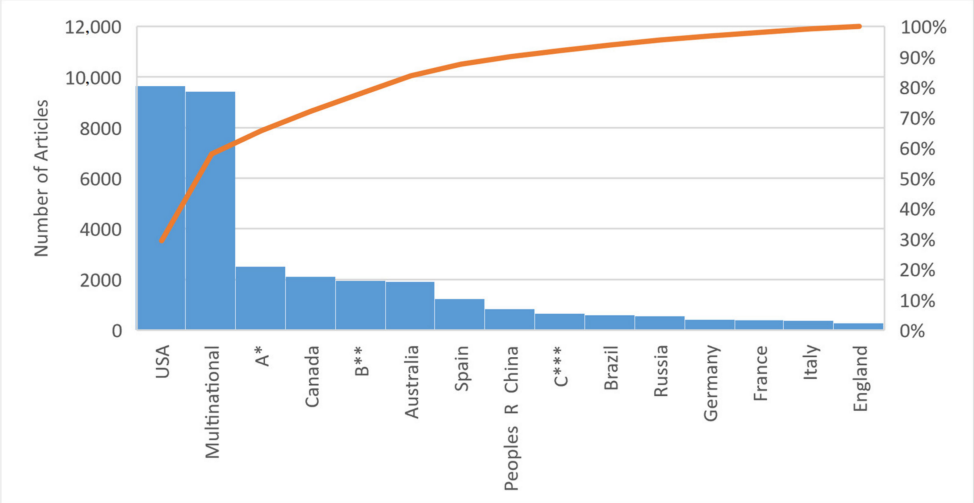

Figure 2. Number of articles by country. A * for countries which have publications between 500 and 999, B ** for countries which have publications between 100 and 499, and C *** for other countries having fewer than 100 publications.

2.2. Text Mining

Text mining is an emerging methodology for semi-automatically extracting information from unstructured text data by indexing words in the text so that relevant information can be extracted [13–15]. It has a set of operations which use text documents as a data source to obtain crucial and structural information. This process requires sophisticated analytical tools that process the text to collect overlooked or important raw data points, along with the necessary keywords. The text mining operations step by step are given as follows:

- Foremost, the database containing the texts related to the researched subject is selected (in this study, the Web of Science database was selected);
- Documents with specific characteristics are filtered by keywords ("forest" or "wild" and "fire", grouped by country);
- Relevant documents from the filtered database are saved in a different format for analysis (documents downloaded as .txt files are saved in MS Excel format and tabulated with many features (title, summary, country, year, number of citations, etc.) to facilitate further analysis);
- Text contents are analyzed with a suitable text mining software (Statistica's text mining tool was used in this study. The use of Statistica's text mining tool along with the user interfaces is given in step-by-step reference [10]);
- The abstracts of the publications were selected as input texts to be mined.

All the abstracts are in English, the indexing language was set as English. The maximum number of words to be chosen was determined as 1000. The minimum frequency of the occurrence of words in a single document to be selected was 3%. Stop words, synonyms, and phrases were defined according to "forest fire" concepts. The stop word list contains words that will not be included in the analysis (e.g., am, is, are, wildfire, forest fire, Elsevier, author, etc.). These words may appear in abstracts, but it is assumed that they do not contain information about the importance of the subject studied. Therefore, it was determined not to be indexed so as to be excluded from the analysis. The synonyms list contains words that should not be evaluated separately (conclude, conclusion; analysis, analyze, analyses; response, responded; south, southern, etc.). The phrase list contains words that should be evaluated as a single word (climate change, remote sensor, etc.).

- Then, words are indexed using the abovementioned stop lists, synonym lists, and phrase lists limited with maximum number of selected words with a minimum occurrence frequency.

Indexed words can be weighted with four different methods that can be used for many kinds of purpose for concept extraction.

1. Raw Frequency: If it is only important how often the words are repeated, the raw statists method, which gives the total number of times the indexed word is repeated in all documents, is used;
2. Binary Frequency: Binary statistics (1 or 0) are used if the use of any word contains valuable information for research;
3. Logarithmic Frequency: Sometimes the repetition of a word in one document more than the other may not mean that the document attaches more importance to that subject at the same rate. In this case, the logarithmic frequency method can be preferred to weight the words:

$$F_i = 1 + \log(wrf_i) \text{ for } wrf_i > 0 \quad (1)$$

where F_i is the logarithmic frequency of indexed word i, and wrf_i stands for row frequency of word i;

4. Inverse Document Frequency (idf):

Inverse document frequency, which includes relative weighting of documents, is the most common method used to weight words within a large number of documents:

$$idf(i,j) = \begin{Bmatrix} 0, & if\ wrf_{ij} = 0 \\ (1 + \log(wrf_{ij})) \log\left(\frac{N}{df_i}\right), & if\ wrf_{ij} \geq 1 \end{Bmatrix} \quad (2)$$

where $idf(i,j)$ refers for inverse document frequency for document j and word i, wrf_i stands for row frequency of word i on document j, N is the total number of documents, and df_i is the document frequency for the i'th word.

In this study, inverse document frequency is used to determine the importance (weights) of selected words and to explore main concepts.

3. Results

The results of this study are given in two subsections as follows: 1-asic text mining results; 2-country-based and year-based results.

3.1. Basic Text Mining Results

In any textual analysis, the repetitions of words may not mean the importance of those words. As the reasons are explained in the sections above, various rules were applied to index words in studies dealing with forest fires and the importance levels of the indexed words were calculated using the inverse document frequency method as in Equation (2).

Figure 3 presents the summary of indexed words in a word cloud scheme and the sizes of words are sized according to their importance weights.

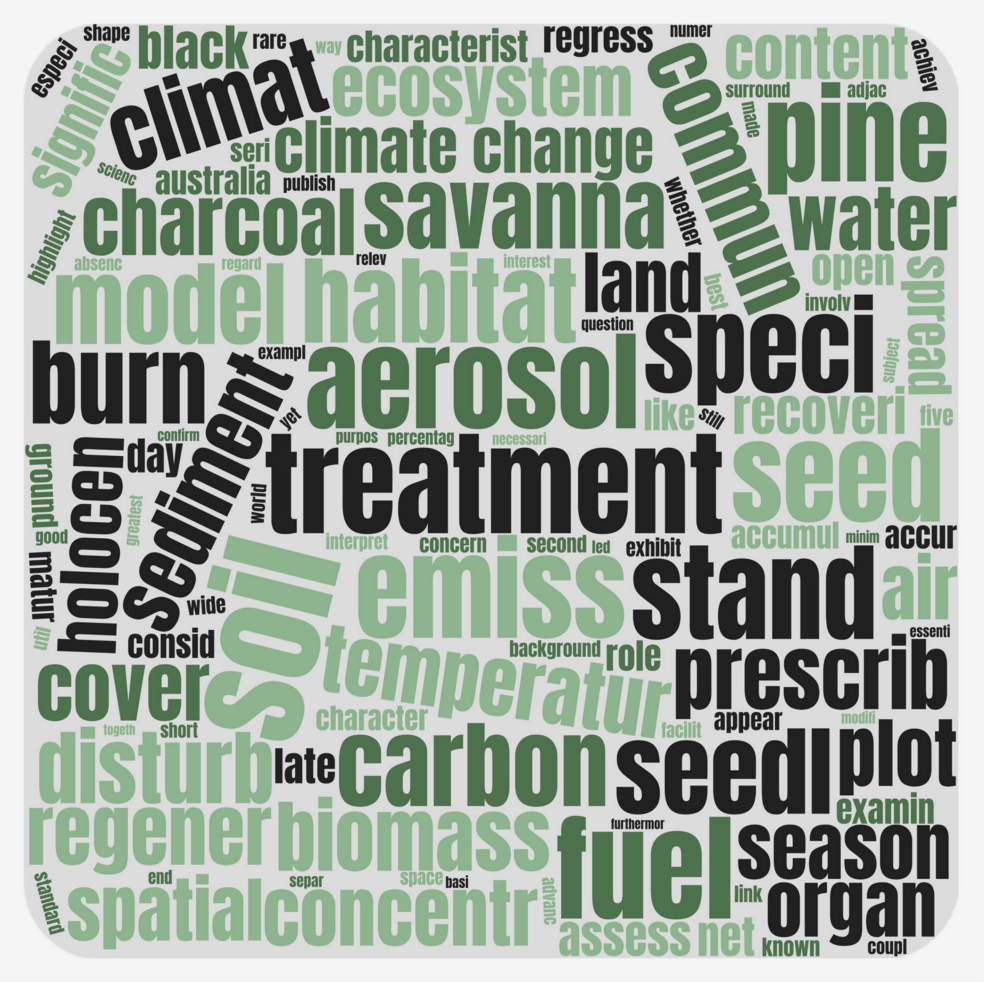

Figure 3. Word cloud of indexed words considering importance weights.

Although each indexed word has a weight of importance, clustering analysis can be used as the most appropriate approach when it is desired to examine which words are the most important. Clustering is the process of separating or grouping a given set of patterns into discrete sets. This is conducted so that patterns in the same cluster are similar and patterns belonging to different clusters are different. In summary, with clustering, similar data can be collected and analyzed in certain groups. Although there are different clustering algorithms, the k-means method has been shown to be effective in producing good clustering results for many practical applications. In this study, it was used in order to analyze the words whose importance weights are close to each other in the same cluster. For the k-means algorithm, the number of clusters (k) must be known beforehand. While performing cluster analysis in Statistica, if the number of clusters is not known, 10-fold analysis can be used instead to find the best number of clusters in certain intervals. The most appropriate number of clusters was searched between 2 and 10 clusters using the k-means clustering method with 10-fold analysis and the words were divided into four clusters according to their importance weights. In Table 6, information on which cluster the indexed words are in is given in descending order of importance weights. The words in the first cluster were found to be the most important words for all time (between 1990 and 2022) in studies on forest fire. It can be said that the studies were mostly on the subjects of soil, fuel, treatment, emission, aerosol, stand, seed, pine, species, tree, oak, seed settlement, habitat, carbon, burn, model, savanna, and climate.

Table 6. Clusters' memberships of indexed words via importance weights.

Clusters	Indexed Words (Descending Ordered by Importance Weights)
Cluster 1	Soil, fuel, treatment, emission, aerosol, stand, seed, pine, species, tree, oak, seed settlement, habitat, carbon, burn, model, savanna, climate
Cluster 2	Disturb, plant, prescribe, biomass, communicate, post-fire, manage, harvestvegetation, regeneration, simulate, plot, temperature, charcoal, sediment, concentrate, cover, season, landscape, water, organ, spatial, site, smoke, growth, rich, wood, air, map, pollen, boreal, sever, ha, year, land, heat, change, composite, population, estimation, area, divers, mortal, risk, ecosystem, thin, Holocene, log, restore, image, density, structure, predict, moisturecanopy, erosion, spruce, lake, data, abundance, increase, product, regime, particle, record, detect, region, CO_2, pollution, cm, class, effect, scenario, surface, climate change, drive, content, atmosphere, tropic, sample, rate, patternconserve, dry, degree, unburn, period, response, age, active, nutrient, survive, differ, size, height
Cluster 3	Measure, pinus, human, global, native, reduce, method, impact, source, condition, signific, layer, accuracy, observe, dynamic, large, total, distribute, control, future, satellite, index, analysis, dominate, low, combust, factor, load, across, time, high, flux, burnt, system, nature, scale, type, recovery, understory, event, diameter, develop, value, spread, transport, weather, black, algorithm, propose, matter, depth, property, Quercus, ecology, influence, stem, relate, loss, annual, litter, precipitate, range, decrease, higher, recruit, frequency, patch, shrub, indicate, base, grass, study, quality, compare, show, import, ignite, process, resolution, rainfall, miner, summer, invasion, suggest, assess, affect, approach, deforest, occurrence, establish, environment, potential, grassland, variate, project, function, net, follow, state, history, provide, Mediterranean, inform, warm, understand, number, occur, reconstruct, unit, result, monitor, biodiverse, level, basal, crown, average, relationship, mass, reduction, correlation, associate, similar, interact, south, mean, ratio, nitrogen, perform, network, open, available, small, intense, radiation, local, month, protect, include, flow, present, decline, early, cause, however, found, century, agriculture, interval, last, late, temporary, timber, current, improv, drought, resource, aim, plan, success, evaluate, research, main, cycle, posit, day, western, locate, degrade, past, policy, behavior, km, nation, determine, deposit, resilience, identify, disperse, mountain, probably, chemic, slope, shift, health, trend, select, reserve, wind, contribute, test, dead, energy, conifer, northern, strategi, ca, need, recent, characteristic, individual, practice, role, anthropogeny, like, respect, regress, strong, suppress, since, experience, long-term, north, first, limit, elevation, field, mechanism, frequent, apply, group, root, negative, examine
Cluster 4	Fragment, extreme, application, remove, new, Canada, right, evidence, variable, live, gradient, well, collect, obtain, cost, combine, eastern, wet, parameter, hazard, service, investigate, fraction, major, spring, Australia, economy, remote sensing, sustain, consider, support, require, succession, winter, damage, distance, maintain, derive, accrue, additive, adapt, date, zone, complex, produce, competition, accumulate, grow, core, material, old, mitigate, mix, fine, experiment, normal, environ, decade, operation, remain, primary, mature, inventory, ground, heterogeny, quantify, explain, initial, amount, case, general, moderate, survey, implement, find, release, integral, part, overall, represent, particular, public, extent, valid, statist, calculation, expanse, enhance, conduct, park, direct, long, character, central, dependent, generate, daily, common, random, benefit, consistent, problem, persist, highest, extract, design, set, stage, previous, tool, uncertainty, maximum, USA, work, decision, prevent, America, continue, account, sensitivity, technique, vulnerable, component, demonstrate, California, key, serial, efficient, transit, dataset, focus, better, form, yield, feature, multiple, reveal, order, basin, emerge, near, element, framework, expect, presence, create, threat, rapid, proport, challenge, knowledge, reflect, specify, critic, suitable, conclude, input, peak, country, lead, terrestrial, contrast, little, action, clear, whereas, report, describe, origin, possible, appear, effort, light, term, even, point, especial, promote, balance, phase, substantial, threaten, return, play, recovery, physic, take, consequent, five, power, close, attribute, format, whether, known, geography, therefor, help, approximate, extensa, address, make, context, contain, capacity, immediate, highlight, achieve, program, experiment, reach, European, vary, second, linear, altern, concern, wide, goal, ability, exhibit, subsequent, aspect, rare, magnitude, link, document, single, carry, shape, despite, discuss, upper, widespread, background, typic, prior, final, issue, least, involve, explore, toward, biology, implicit, distinct, lack, correspond, poor, standard, fire-prone, utility, either, publish, incorporate, couple, great, consider, world, interpret, surround, define, facility, short, hypothesis, space, adjacent, place, percentage, detail, advance, numerical, independent, good, way, example, absence, led, relevance, regard, purpose, question, greatest, science, except, subject, complete, interest, separate, basic, necessary, minimum, essential, modify, confirm

Just as many comments can be made from the indexed words and their importance weights, another function of text mining is to extract concepts. Instead of examining all words one by one, all indexed words can be converted into fewer variables with singular value decomposition (SVD) and analyzed. SVD is a method of converting related variables into a set of unrelated variables that better reveal the various relationships between the original data items. In addition, SVD can be used to identify and rank the dimensions for which data points show the greatest variation. After determining where there is the greatest variation, it is possible to find the best approximation to the original data [16].

This approach is also referred to as concept extraction. Figure 4 shows that the indexed words can be analyzed according to their importance weights in twenty-nine concepts. As can be seen from the Figure 4, each concept has a different weight. While the first concept weighs considerably more than the others, the second, third, and fourth concepts have relatively greater weights than the others. In addition to these, the weights of the remaining concepts are both decreasing and are very close to each other.

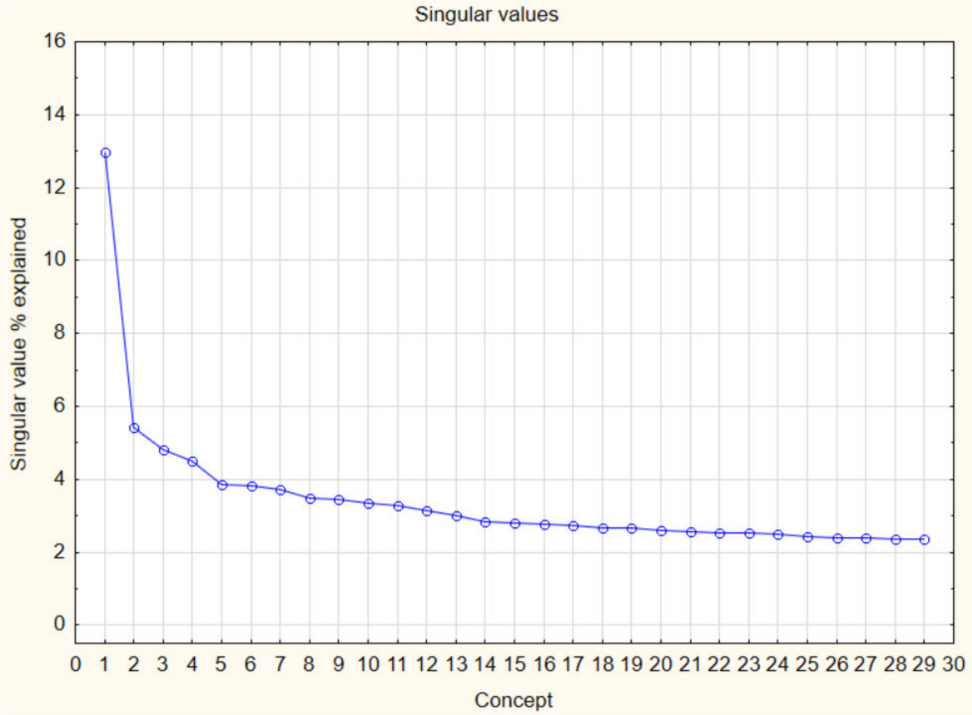

Figure 4. SVD value percentages of concepts.

As with the indexed words, the resulting concepts were also clustered by weight with k-means 10-fold analysis and divided into five clusters as seen in Table 7. Concepts 1 and 2 were clustered alone, while Concepts 3 and 4 were clustered together. In addition, Concepts 5 to 12 and Concepts 13 to 29 were clustered together separately. Each indexed word has different weights in the concepts. Table 8 shows the 10 words with the most significant weight for each concept.

Table 7. Class memberships of extracted concepts by singular value decomposition via weights.

Class	Concept
1	Concept 1
2	Concept 2
3	Concepts 3–4
4	Concepts 5–12
5	Concepts 13–29

Table 8. Top ten indexed words via importance weights for each concept.

Concept ID	Top Ten Indexed Words (Ordered by Importance Weight)	Concept ID	Top Ten Indexed Words (Ordered by Importance Weights)
Concept 1	Species, burn, soil, tree, model, vegetation, area, site, year, change	Concept 16	Seed, fuel, spruce, stand, boreal, regeneration, emission, temperature, moisture, rich
Concept 2	Species, tree, stand, seed settlement, regeneration, pine, treatment, seed, plant, plot	Concept 17	Erosion, cover, sediment, log, canopy, water, aerosol, surface, flow, rainfall
Concept 3	Manage, map, landscape, risk, model, spatial, approach, habitat, inform, plan	Concept 18	Habitat, simulate, harvest, water, landscape, seed, restore, soil, temperature, treatment
Concept 4	Pollen, Holocene, charcoal, climate, record, lake, year, sediment, reconstruct, human	Concept 19	Heat, detect, image, fuel, smoke, log, boreal, temperature, surface, degree
Concept 5	Soil, communicate, ecosystem, organ, carbon, conserve, plant, nutrient, biodiverse, manage	Concept 20	Pine, habitat, oak, unburn, species, mortal, population, climate change, pinus, emission
Concept 6	Emission, carbon, fuel, climate change, CO_2, treatment, climate, future, stand, scenario	Concept 21	Treatment, heat, degree, aerosol, simulate, scenario, size, thin, particle, ha
Concept 7	Aerosol, emission, species, smoke, habitat, air, pollution, rich, source, atmosphere	Concept 22	Treatment, thin, map, mortal, lake, climate, charcoal, class, tree, density
Concept 8	Carbon, boreal, soil, disturb, map, CO_2, spatial, net, ecosystem, global	Concept 23	Pine, flux, heat, temperature, CO_2, scale, water, flow, spatial, pinus
Concept 9	Stand, charcoal, year, harvest, carbon, ha, pollen, sediment, lake, age	Concept 24	Seed, restore, drive, ha, dry, resilience, sample, range, tree, ecology
Concept 10	Fuel, burn, prescribe, rich, habitat, divers, load, cover, treatment, communicate	Concept 25	Oak, post-fire, burn, aerosol, charcoal, Quercus, heat, unburn, manage, height
Concept 11	Seed settlement, burn, season, image, treatment, cover, land, satellite, year, annual	Concept 26	Oak, harvest, wood, heat, product, plant, combust, concentrate, Quercus, class
Concept 12	Stand, post-fire, sever, log, concentrate, event, boreal, spruce, harvest, erosion	Concept 27	Plot, cm, depth, communicate, ha, soil, layer, air, nation, surface
Concept 13	Ha, population, tree, cm, diameter, protect, conserve, savanna, risk, habitat	Concept 28	Savanna, emission, drive, air, pollution, treatment, tropic, propose, CO_2, spruce
Concept 14	Seed, post-fire, sediment, erosion, patch, fuel, landscape, disperse, simulate, burnt	Concept 29	Spruce, risk, boreal, black, zone, slope, cost, energy, cover, communicate
Concept 15	Harvest, habitat, rich, abundance, treatment, log, rate, response, species, number		

Figure 5 is given in order to give an idea of the importance weights of the words given for each concept and what topics they cover. In the word clouds given in this way, the sizes of the words are in line with the ratio of importance weights in that concept. The larger word means more importance, while the smaller font word has less importance weight.

Figure 5. Word clouds of each concept for ten most important words.

3.2. Country and Year Based Results

In this section, various graphs are presented to understand how the importance weights of some selected indexed words and concepts behave over the years and countries. After the word indexing process, more than five hundred words were indexed and only some words were mentioned in this study. These words have been chosen from words that have changed in importance weights over the years and will cover current issues, and they are discussed one by one below. In Figure 6, there is a heatmap representation of the differences in importance weights of these words by country. In addition, the numbers in the figure show the means of importance weights of the given word for the given country. A higher weight means that the word is more important to that country than others, while

a lower weight means less importance. Figure 7 shows the change in the means of the importance weights of the selected words over the years.

Countries / Words	Fuel	Soil	Treatment	Emission	Carbon	Climate Change	Remote Sensing	Management	Model	Post-fire
A	45	51	24	36	35	23	14	51	52	27
Australia	71	65	44	20	36	29	9	71	58	43
B	26	71	20	32	34	21	14	40	49	34
Brazil	11	80	14	33	33	14	18	30	42	23
C	22	77	20	21	24	22	18	45	47	14
Canada	31	51	30	20	33	36	12	64	63	40
England	33	19	5	53	34	28	13	33	67	12
France	70	50	16	36	30	16	5	26	75	13
Germany	20	57	18	44	37	26	21	22	64	3
Italy	41	81	13	28	18	17	25	44	59	40
Multinational	38	71	24	58	63	34	16	44	67	30
Peoples_R_China	34	61	16	52	51	22	22	28	84	17
Russia	21	84	2	48	63	12	11	6	40	19
Spain	42	116	63	14	21	17	19	49	56	72
USA	69	57	78	28	39	28	11	72	60	34

Figure 6. Heatmap of word mean importance weights ($\times 10^{-2}$) by countries.

Years / Words	Fuel	Soil	Treatment	Emission	Carbon	Climate Change	Remote Sensing	Management	Model	Post-fire
Year 1990	70	94	0	0	0	0	0	14	16	16
Year 1991	55	32	33	7	19	12	6	19	27	20
Year 1992	71	22	16	29	27	5	0	22	43	8
Year 1993	86	25	34	25	21	10	8	23	47	18
Year 1994	79	28	31	20	15	8	7	37	29	11
Year 1995	69	29	26	28	30	5	9	28	39	14
Year 1996	72	23	25	29	23	4	3	33	50	14
Year 1997	70	26	26	18	28	7	9	33	43	18
Year 1998	79	29	30	26	27	13	1	35	42	18
Year 1999	66	22	34	13	23	5	5	32	50	26
Year 2000	76	29	27	22	40	10	11	42	47	20
Year 2001	69	39	30	25	25	9	7	41	40	20
Year 2002	61	30	30	32	48	8	12	47	48	23
Year 2003	65	38	39	32	37	7	17	43	45	17
Year 2004	74	45	44	30	37	8	10	44	53	22
Year 2005	83	42	49	28	38	10	10	49	46	26
Year 2006	69	39	57	34	35	12	9	46	50	31
Year 2007	67	53	58	33	37	15	9	53	60	21
Year 2008	67	46	59	31	37	16	12	50	52	27
Year 2009	67	52	53	43	47	23	17	52	58	21
Year 2010	70	49	60	34	41	26	12	52	53	36
Year 2011	61	52	39	48	54	29	14	49	63	28
Year 2012	69	59	49	45	44	27	9	55	60	32
Year 2013	65	56	47	48	50	30	10	58	61	34
Year 2014	55	66	44	44	48	29	13	53	65	34
Year 2015	66	49	41	46	49	37	11	58	64	33
Year 2016	68	52	43	42	54	37	16	62	64	47
Year 2017	67	53	38	37	51	36	11	58	63	33
Year 2018	61	52	44	42	50	36	18	58	64	35
Year 2019	61	45	37	38	45	37	14	61	63	40
Year 2020	62	46	40	38	47	40	18	62	68	42
Year 2021	57	47	33	34	45	39	21	60	71	38
Year 2022	61	52	30	40	46	42	26	62	77	48

Figure 7. Heatmap of word mean importance weights ($\times 10^{-2}$) by year.

Word by word analysis results are given in Figures 8–12. In each Figure, only two words' graphics are given in order not to disturb the page layout in any way. Average graphs were drawn for every five years in order to determine the change in the importance level of the related words over the years. Analysis of variance graphs are provided to observe the mean weights of importance of the relevant terms for each country at all times. For all the results obtained, p values less than 0.05 were reached in the 0.95 confidence interval, and it was determined that the differences in the mean weights of importance

between years and countries were significant. (a) and (b) terms in figures represent the first two items, 1 represents the change according to years, and 2 represents the change according to the countries.

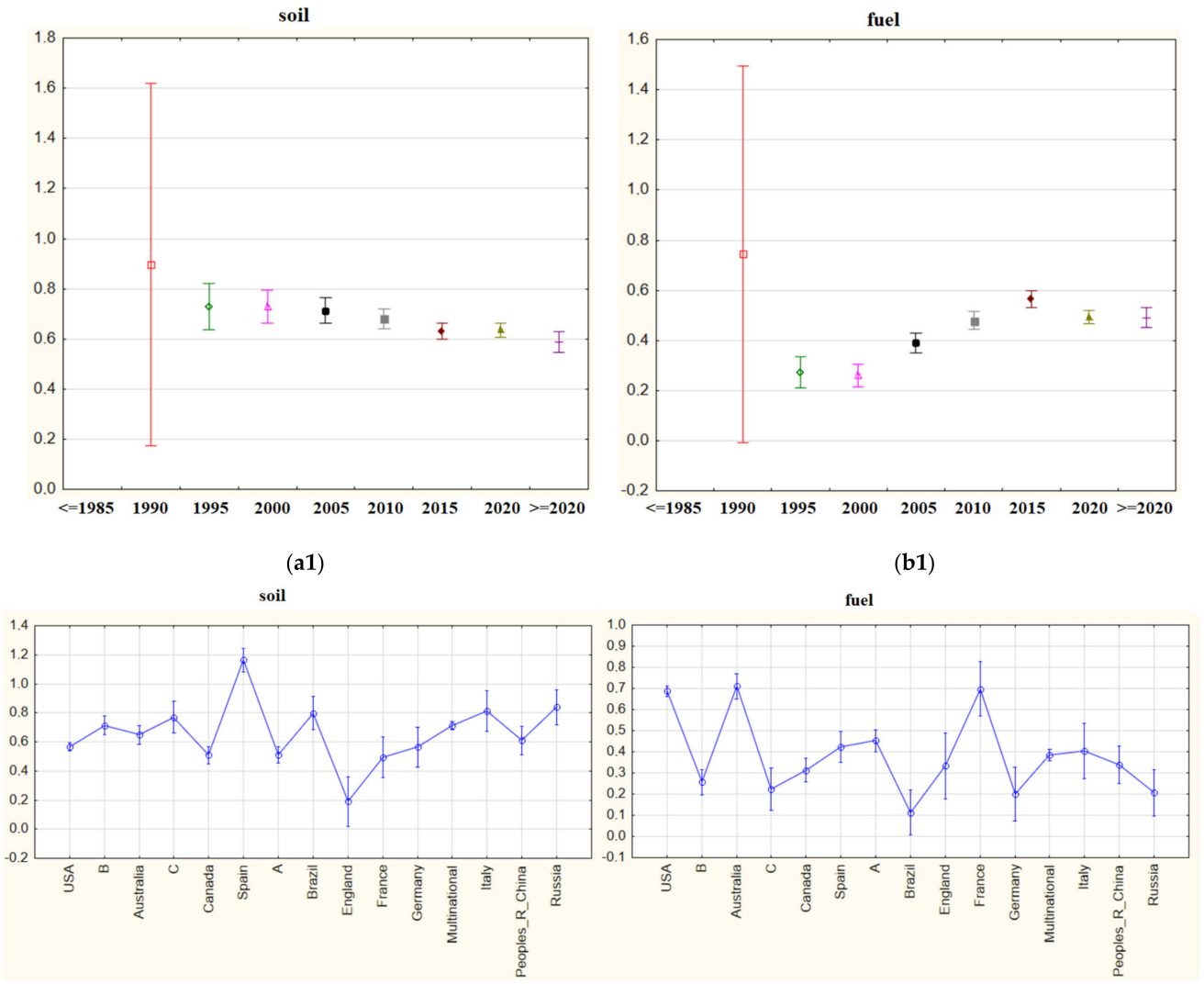

Figure 8. (**a1**) mean importance weights of the indexed word "soil" for five-year periods; (**b1**) mean importance weights of the indexed word "fuel" for five-year periods; (**a2**) mean importance weights of the indexed word "soil" via countries; (**b2**) mean importance weights of the indexed word "fuel" via countries.

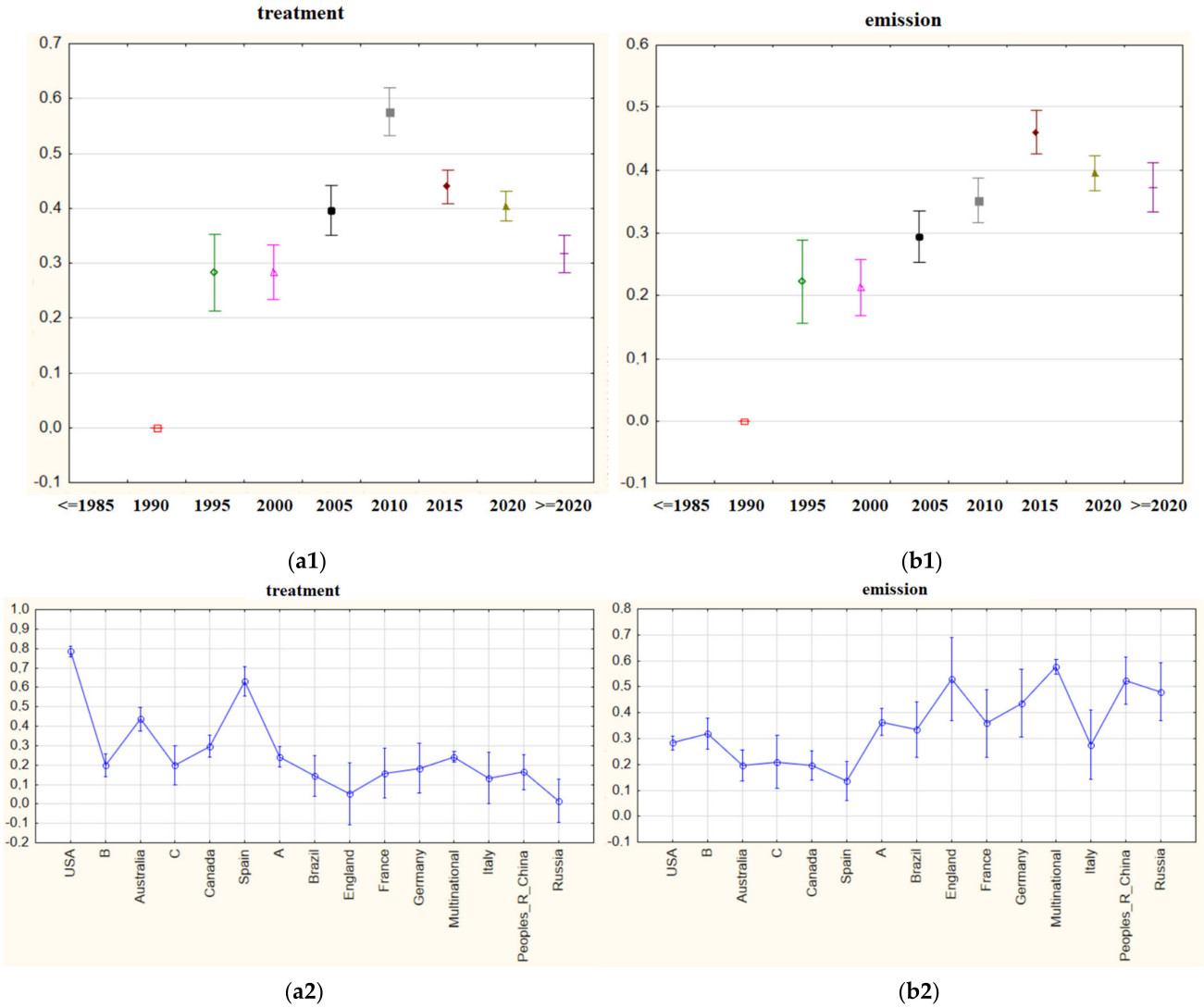

Figure 9. (**a1**) mean importance weights of the indexed word "treatment" for five-year periods; (**b1**) mean importance weights of the indexed word "emission" for five-year periods; (**a2**) mean importance weights of the indexed word "treatment" via countries; (**b2**) mean importance weights of the indexed word "emission" via countries.

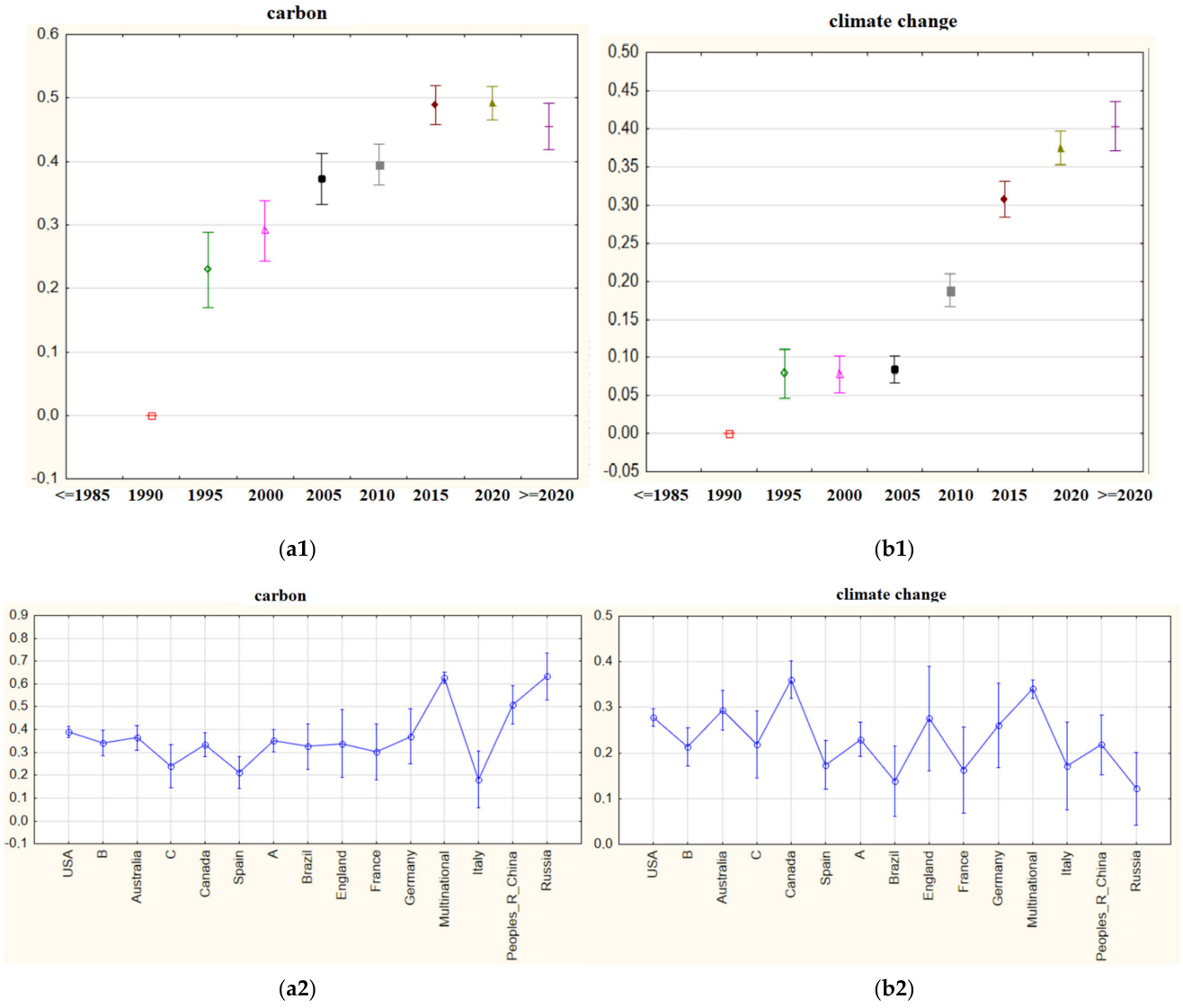

Figure 10. (**a1**) mean importance weights of the indexed word "carbon" for five-year periods; (**b1**) mean importance weights of the indexed term "climate change" for five-year periods; (**a2**) mean importance weights of the indexed word "carbon" via countries; (**b2**) mean importance weights of the indexed term "climate change" via countries.

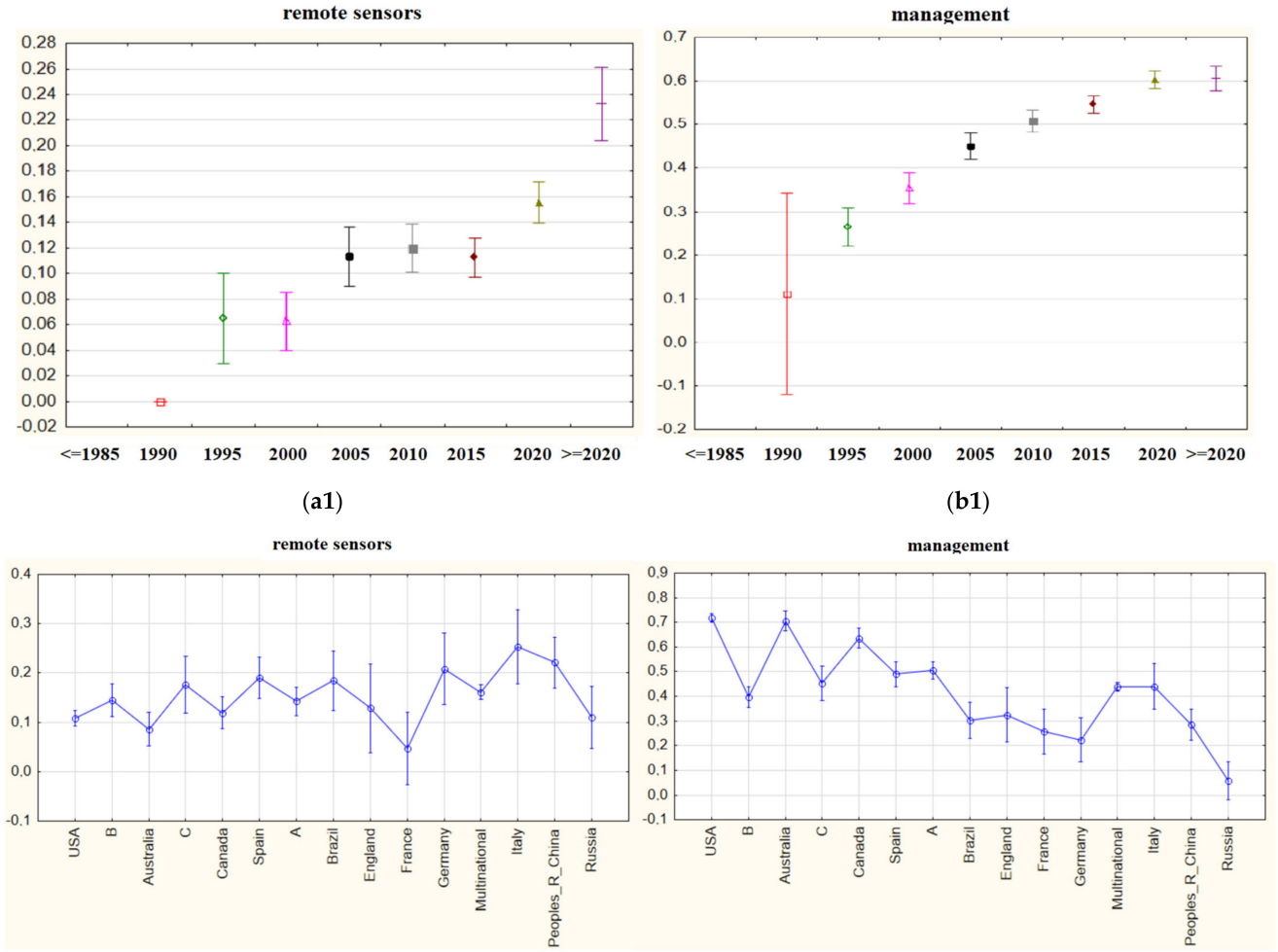

Figure 11. (**a1**) mean importance weights of the indexed term "remote sensors" for five-year periods; (**b1**) mean importance weights of the indexed word "management" for five-year periods; (**a2**) mean importance weights of the indexed term "remote sensors" via countries; (**b2**) mean importance weights of the indexed word "management" via countries.

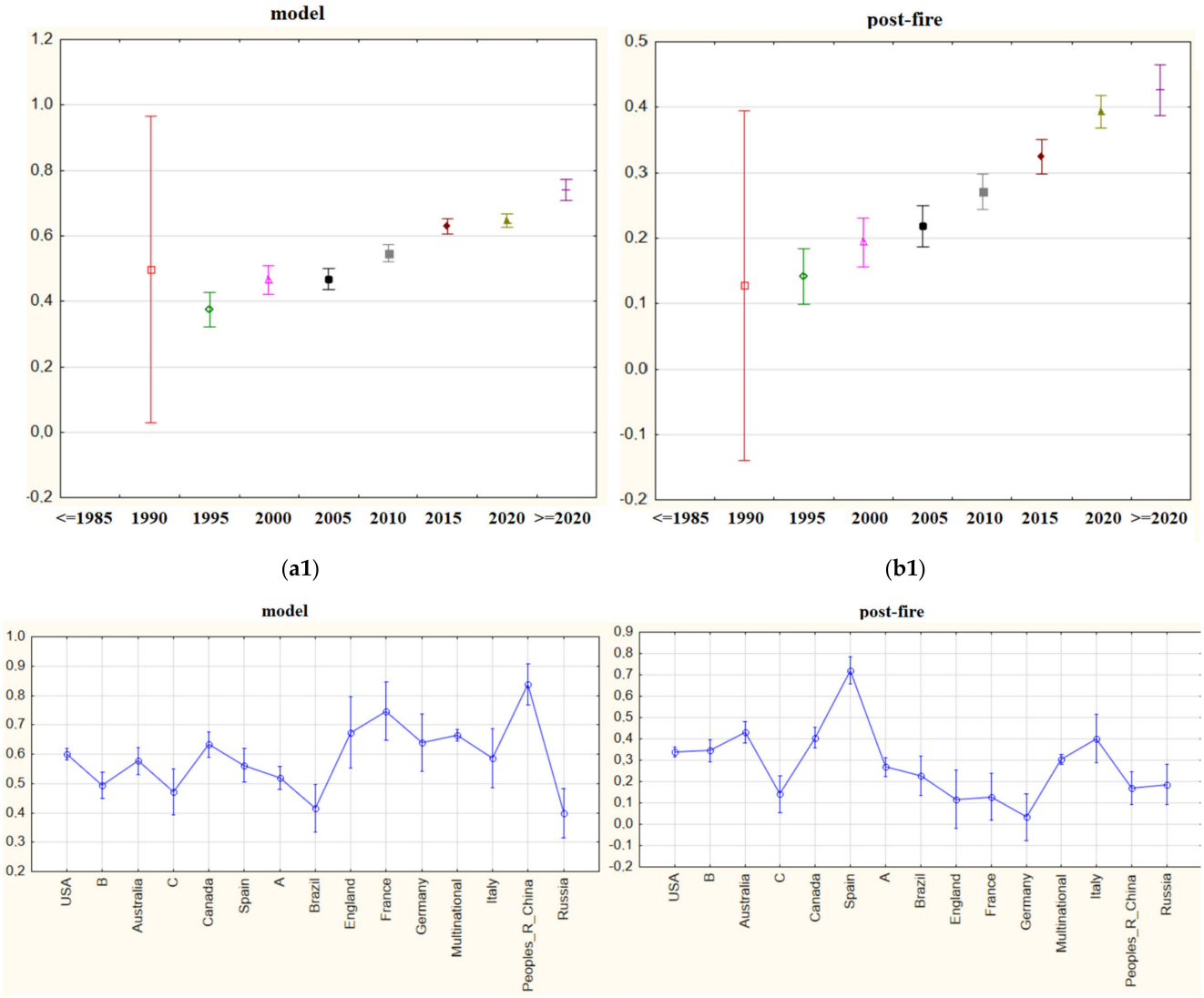

Figure 12. (**a1**) mean importance weights of the indexed word "model" for five-year periods; (**b1**) mean importance weights of the indexed term "post-fire" for five-year periods; (**a2**) mean importance weights of the indexed word "model" via countries; (**b2**) mean importance weights of the indexed term "post-fire" via countries.

"soil" and "fuel" were the first two words with the greatest importance, respectively. In Figure 8, the differences in the mean of importance weights of these two words according to years and countries can be observed.

The most important word in forest fire studies was calculated as the word "soil". This may actually mean that the subject of how the soil is affected by forest fires is being investigated. However, over each five-year period, the word has neither gained much importance nor lost it significantly. It has been observed that while Spain gives the highest importance to "soil" in forest fire studies, England gives the least importance.

The second most important word was "fuel". Contrary to the word "soil", the word "fuel" can be said to be a subject whose importance has increased over the years with a slightly increasing trend after the year 2000. What is meant by the word "fuel" should generally be perceived as combustible organic materials (such as dried leaves and dried wood residues) accumulated in forested areas. The word may encompass research that combines some sort of accumulated biomass and probable wildfire research. It is observed that the USA, Australia, and France are the countries that attach the most importance to

research on the relationship of combustible material accumulated in forest areas with forest fires. On the other hand, studies originating in Brazil, Germany, and group C countries gave less importance than the others.

"treatment" and "emission" were the third and fourth biggest importance weighted indexed words, respectively. Figure 9 shows the differences in the mean of importance weights of these two words according to years and countries.

The word "treatment" is a topic that covers the treatment of the region after the forest fire. It is observed that the word, which had an increasing trend until 2010, lost its importance in the following years. It is very clear from Figure 9(a1) that the subject of "treatment" is a topic that has lost its importance for forest fire research, and this can be perceived as a sign that new research plans should focus on different subjects. Although the USA is the country that attaches the most importance to this issue, other countries, except for Australia and Spain, did not assign much importance to this area.

The word "emission" is another important word. We can see that this word is used frequently, especially in studies dealing with gas emissions released after a fire. This is another post-fire related research topic which is important. This issue had a clear increasing trend until 2015 then started to lose its importance, changing to a decreasing trend. It has been observed that research with multinational addresses attach the most importance to this issue. In addition, England was another country that gave higher importance than most countries.

The word "carbon" may also be related to emissions as well as the amount of carbon in combustible materials accumulated in forest areas. It can be observed in Figure 10 that it has an increasing trend, although it has a horizontal importance in the last fifteen years. As an interesting finding, studies from Russia and multinational addresses are leading in studies that give high importance to "carbon", whereas the weight of importance for other countries is relatively low and similar.

Especially after the Kyoto protocol, which entered into force on 16 February 2015, many countries gave importance to climate change in many different areas. As can be seen in Figure 10(b1), it can be observed that the importance given to climate change has increased significantly in studies on forest fires, especially after 2005. As can be seen from the figure, the positive direction in the trend has become sharply evident and it has not been revealed to have lost its importance yet. It is likely that forest fires and climate change linkage (for pre-fire and post-fire situations) will continue to be investigated in the following periods as well. Research from Canada and multinational addresses gave the greatest importance to this issue; conversely, studies from Russia gave the least importance.

Although the words examined in this section are primarily selected from the first cluster containing the most important words in Table 6, it will not be possible to analyze all the indexed words one by one. Some indexed words that are not among the most important cluster may have a possible positive trend and need to be analyzed. An example of these words is "remote sensing". Remote sensing is an issue that includes the use of various technologies, which have an important place in early fire intervention.

It can be observed in Figure 11 that the phrase demonstrates a remarkable positive trend in the last ten years; however, it has a low importance among all the studies examined and is therefore in the fourth cluster. Researchers from Italy and then China assign the most importance to remote sensing and forest fire studies, and France has the lowest importance level.

Today, "management", which has importance in many different sectors, also demonstrates significant importance in the management of forest fires. It is a subject worth investigating; its level of importance shows a positive trend over the years. The USA and Australia attach immense importance to studies related to forest fire management; Russia, by contrast, assigned it the least importance.

In order to manage anything, it is necessary to have data about it and model it. One of the findings of this study shows that the word "model" has a prominent place among studies on forest fire. As can be seen in Figure 12, the importance given to modeling has

been increasing over the years and is still up to date. China attaches the greatest importance to this issue, whereas Brazil and Russia give relatively lower importance than other.

Interestingly, some of the topics related to pre-fire, during-fire, and post-fire are of increasing importance, while some others are becoming obsolete. Therefore, not-significantly-trending or obsolete issues related to fire timing have been identified. The term "post-fire", however, has emerged as a term with a positive trend on its own. Spain gave the term the greatest importance, while Germany gave it the lowest importance.

In Figure 13, the differences in the average importance weight of each concept by country are shown as a heatmap. Table 8 and Figure 5 contained information about which topics the concepts covered. In Figure 13, it can be easily observed whether a particular country attaches more importance to any concept than others. For example, while Concept 12, which probably focused on studies related to the density of spruce cut after the fire, was mostly found in studies based on the USA and Canada, other countries did not give much importance to this issue in studies related to forest fires. Again, for example, Concept 6, which probably focuses on studies on carbon emissions and climate change, was found to be significantly more important in Canada-based studies, while other countries did not attach much importance to this issue compared to Canada.

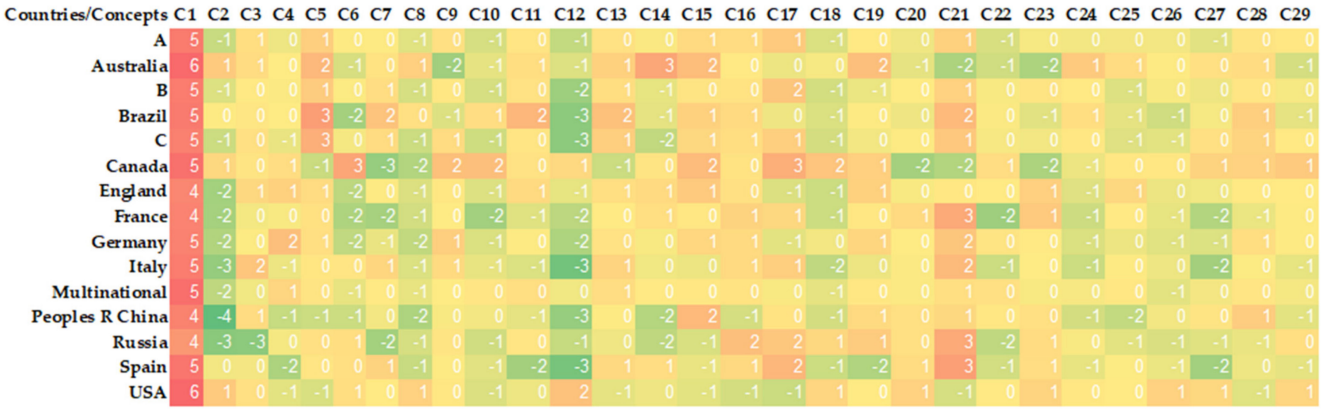

Figure 13. Heatmap of created concepts mean importance weights ($\times 10^{-3}$) by countries.

Figure 14 shows the heatmap based on the means of the importance weights of the concepts. From this figure, one can easily observe trending and de-trending concepts over the years. For example, it can be said that Concept 1 has a trend of increasing importance weight; by contrast, although it has second importance, Concept 2 is de-trending and now these issues are given less importance.

Since it will be difficult and time consuming to examine all indexed words one by one, it is possible to acquire an idea about trending and obsolete topics by looking at the concepts obtained in general. Twenty-nine concepts obtained were analyzed one by one, but it was observed that those whose graphs are given below demonstrate a clear sign of trending or obsolescence.

Some selected concepts' analysis results are given in Figures 15–19. Table 8 and Figure 5 above contain the ten most important words of each concept. These words give clues about the basic contents of the concepts. Concept 1 includes the topics of burned species, trees, areas, and vegetation, while Concept 2 includes the topic of post-fire regeneration of species, trees, areas, and vegetation. As can be seen in Figure 15, Concept 1 has an increasing importance day by day, while Concept 2 is an obsolete topic. The studies prepared by the USA, Australia, and multinational authors gave more importance to Concept 1, and it can be seen that the least importance was assigned by Russia addressed-studies. On the other hand, while the USA, Australia, and Canada gave the most importance to Concept 2, China gave the least importance.

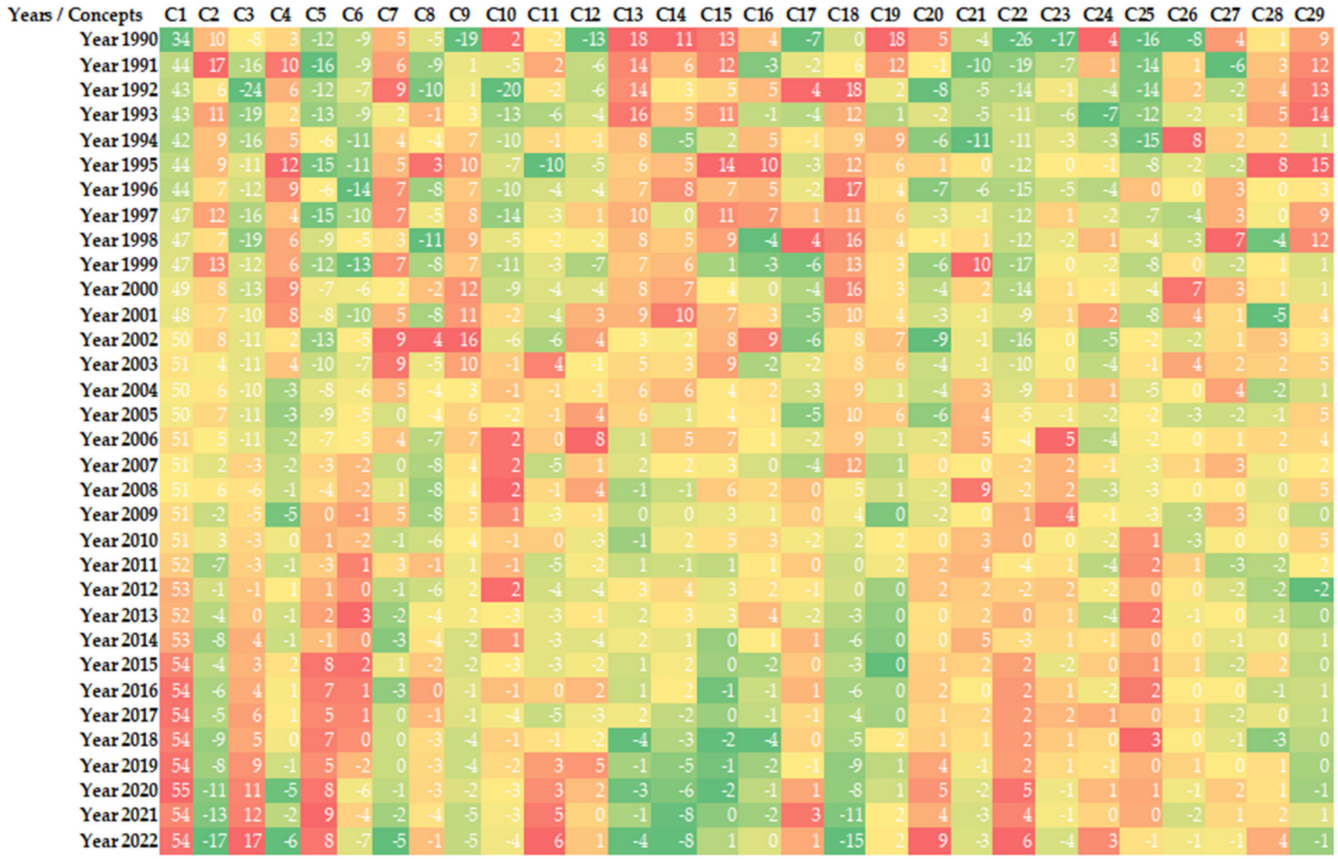

Figure 14. Heatmap of created concepts mean importance weights ($\times 10^{-4}$) by year.

In Figure 16, it is observed that the third concept has a strong positive importance trend. This concept, which includes research topics such as management, risk management, and habitat, is considered up-to-date and worth investigating. The interesting result of the research is that the importance assigned to the subject in articles originating from other countries except Russia is close to each other. The fourth concept, which focuses on geological structures such as the Holocene, aged vegetation, pollen, and sediment, lost serious importance, especially after 2000, and researchers moved away from this area. The researchers who attach the greatest importance to this concept are from Germany.

It can be seen from Figure 17 that Concept 5, which includes ecosystem components, has an increasing trend. Brazilian authors gave the highest emphasis to ecosystem-focused wildfire studies, while Canadians gave the lowest importance. As a very interesting finding, the sixth concept, which includes topics such as the ecosystem, carbon emissions, and climate change, was trending until 2015, but then it started to become obsolete. The authors who valued this concept most were the Americans and Canadians, while Italian and Brazilian authors gave the least value.

In particular, the results of Concept 18 can be examined in Figure 18, giving examples of concepts that have lost their importance. In addition, the graphs given in Figures 18 and 19 for Concepts 20, 22, and 25—which are trending and can be perceived as possible research topics—can be examined, even though their importance weights are not relatively high.

These results contain particularly important clues about the importance and trends of the topics while choosing new research topics. In this study, the systematic approach applied as a case study on forest fires can be easily applied in any field and new study topics can be determined by analyzing the results.

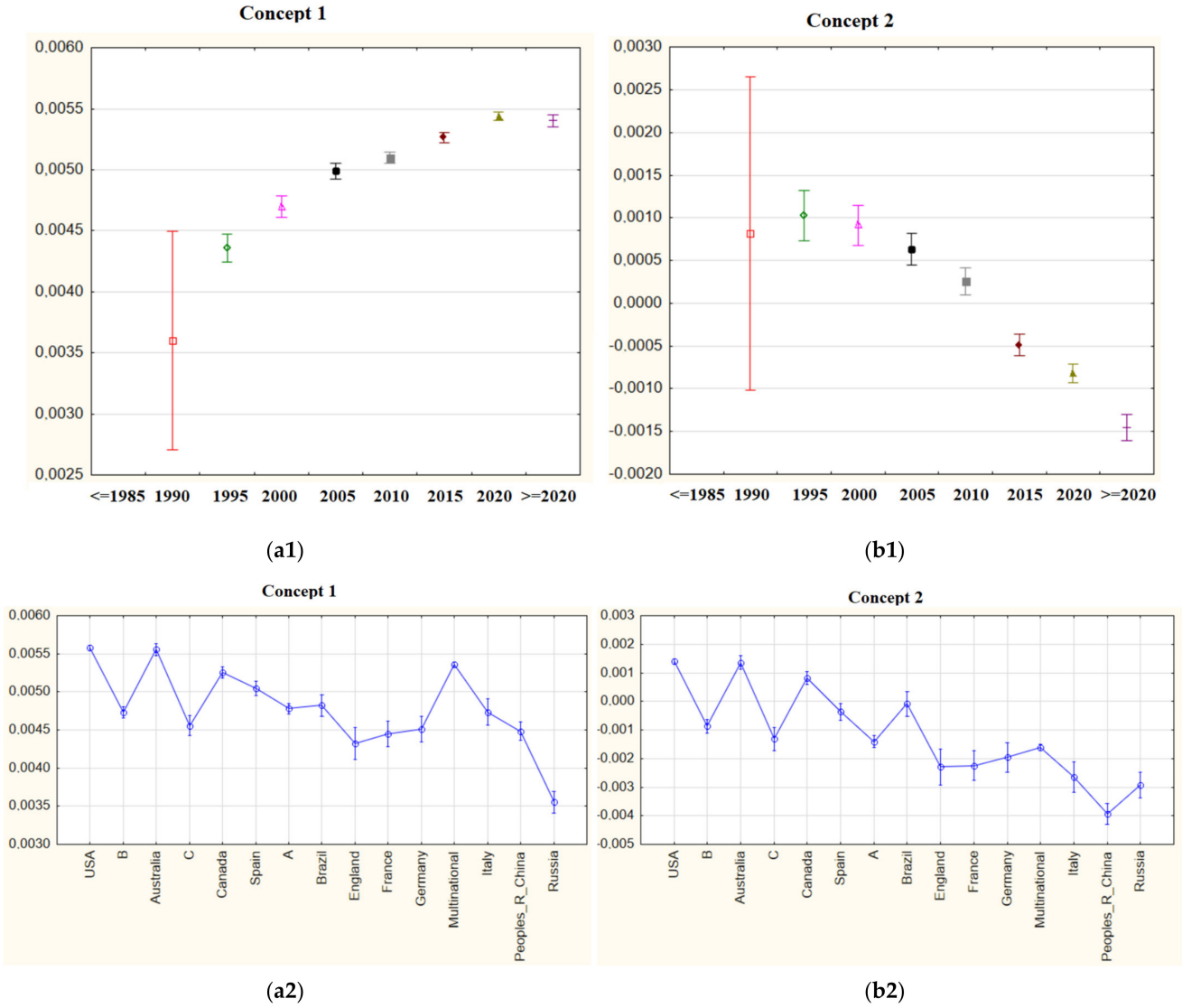

Figure 15. (**a1**) mean importance weights of Concept 1 for five-year periods; (**b1**) mean importance weights of Concept 2 for five-year periods; (**a2**) mean importance weights of Concept 1 via countries; (**b2**) mean importance weights of Concept 2 via countries.

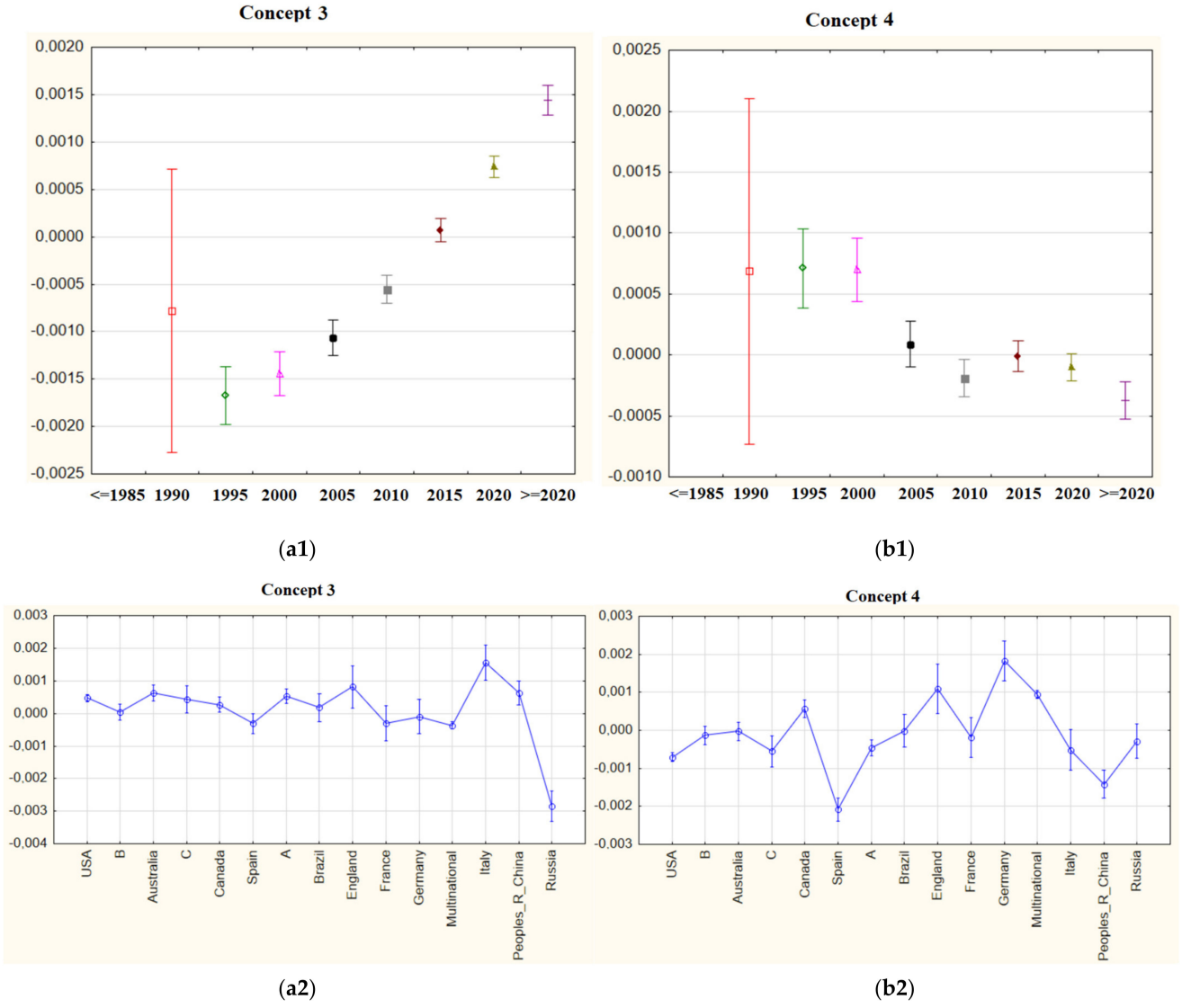

Figure 16. (**a1**) mean importance weights of Concept 3 for five-year periods; (**b1**) mean importance weights of Concept 4 for five-year periods; (**a2**) mean importance weights of Concept 3 via countries; (**b2**) mean importance weights of Concept 4 via countries.

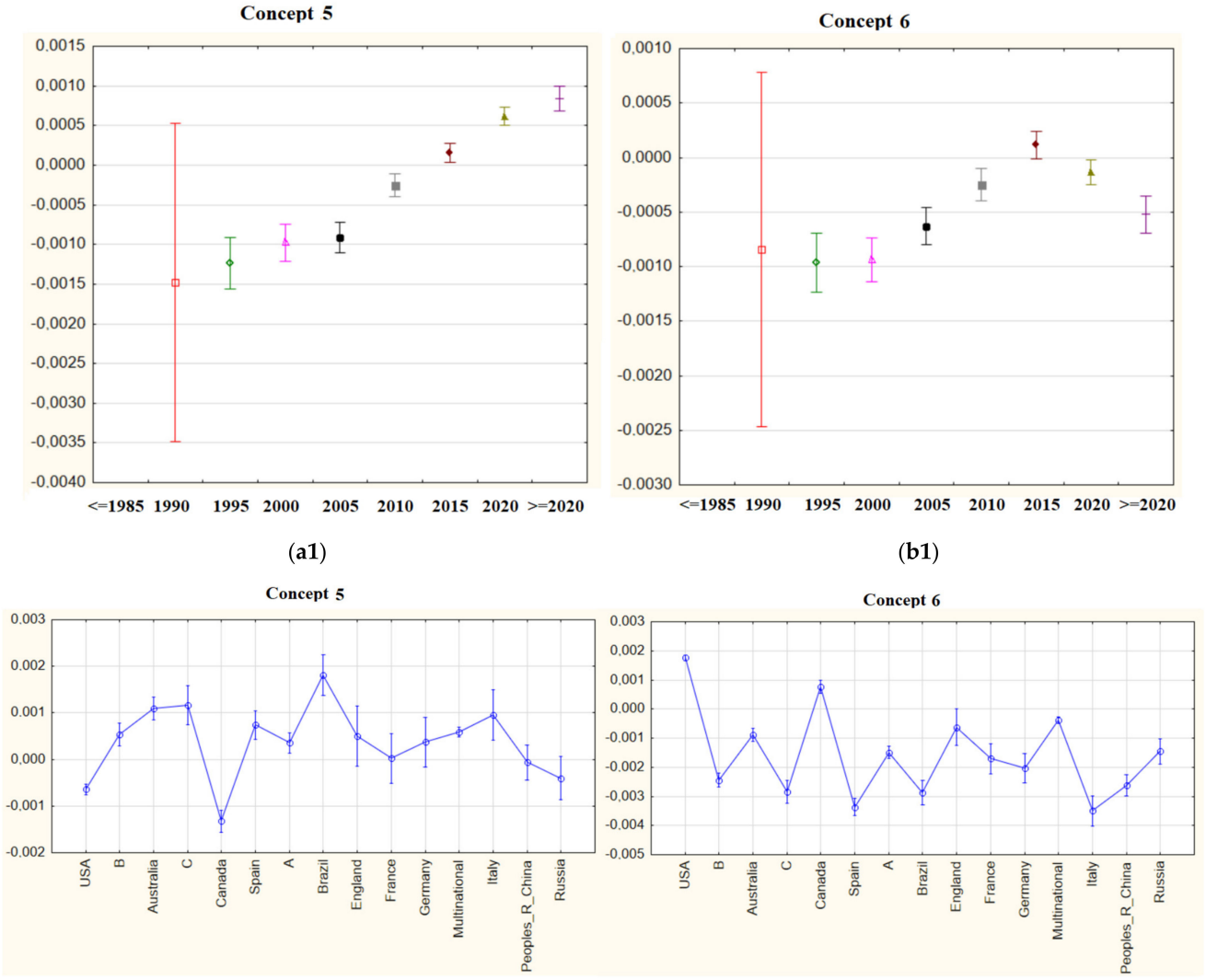

Figure 17. (**a1**) mean importance weights of Concept 5 for five-year periods; (**b1**) mean importance weights of Concept 6 for five-year periods; (**a2**) mean importance weights of Concept 5 via countries; (**b2**) mean importance weights of Concept 6 via countries.

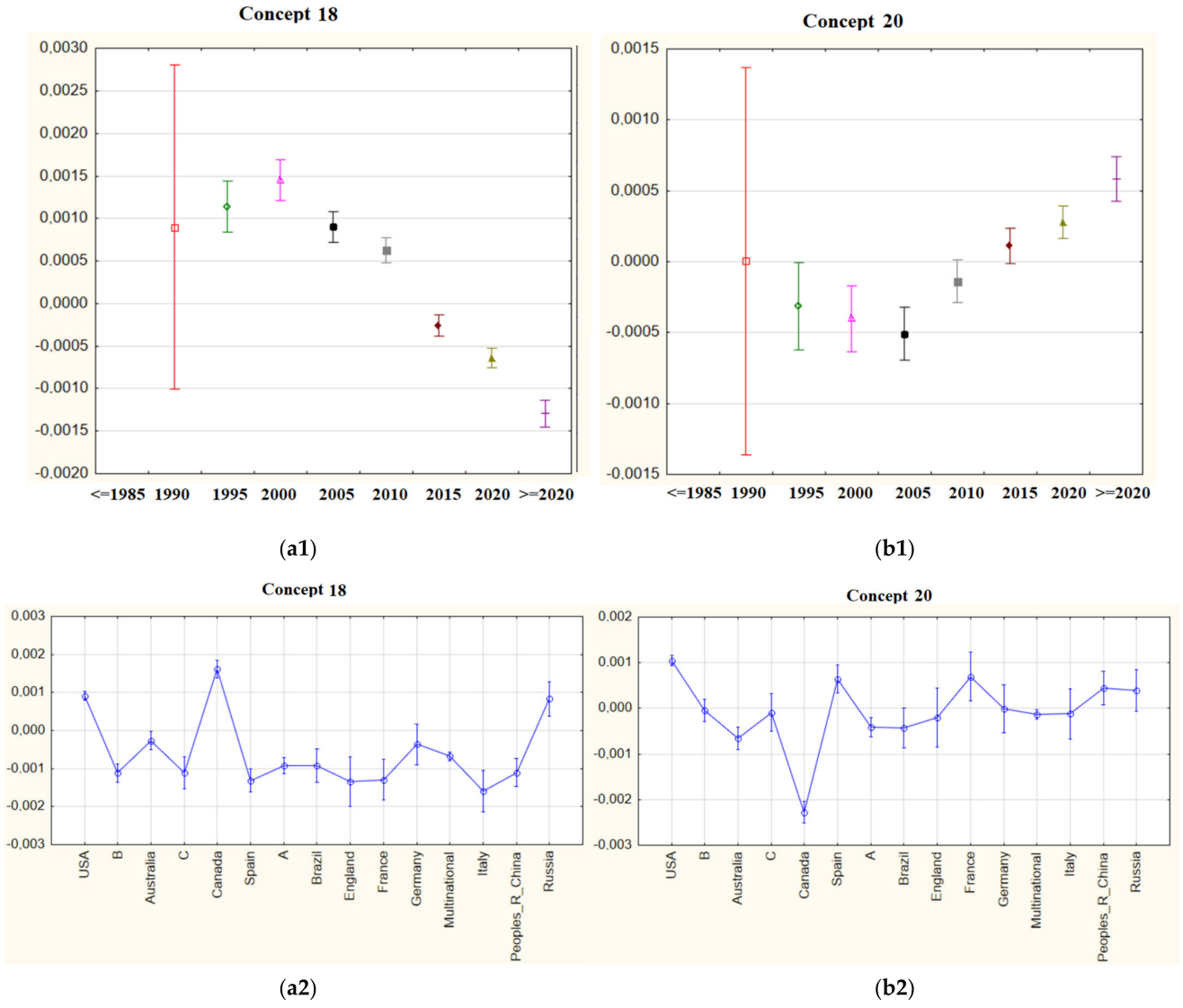

Figure 18. (**a1**) mean importance weights of Concept 18 for five-year periods; (**b1**) mean importance weights of Concept 20 for five-year periods; (**a2**) mean importance weights of Concept 18 via countries; (**b2**) mean importance weights of Concept 20 via countries.

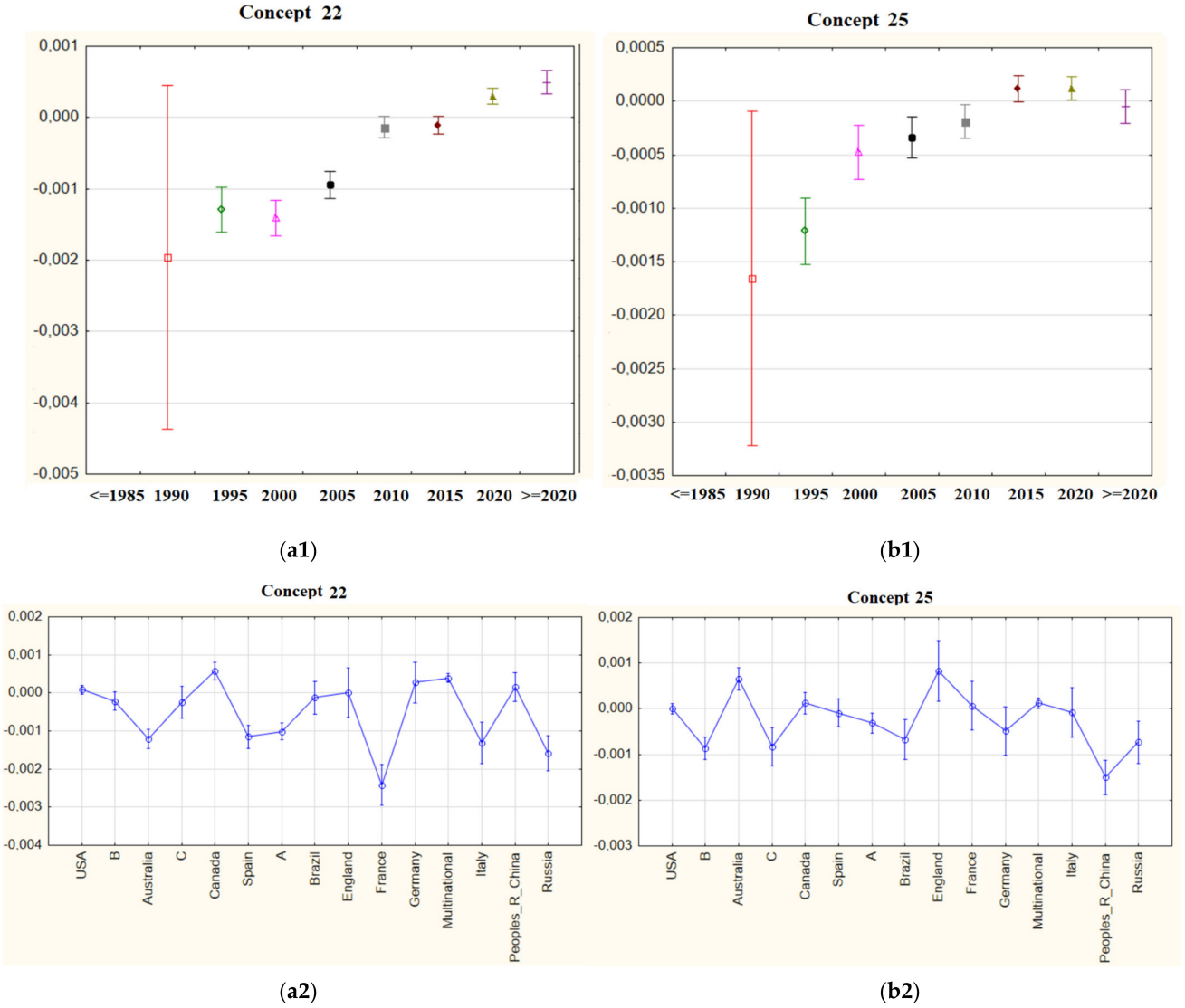

Figure 19. (**a1**) mean importance weights of Concept 22 for five-year periods; (**b1**) mean importance weights of Concept 25 for five-year periods; (**a2**) mean importance weights of Concept 22 via countries; (**b2**) mean importance weights of Concept 25 via countries.

4. Discussion and Conclusions

Academic studies on forest fires in the Web of Science database, which is accepted by the scientific community, were collected in a specific structure, and analyzed by text mining. The main contribution of the study to the literature is that it represents a systematic approach to identify trending and obsolete topics for the purpose of guiding new studies in a particular field. Although there are studies on concept extraction and trend setting in different fields such as energy or aviation in the current literature, no studies focused on forest fire have been found.

In this study, word indexing was first conducted within the framework of certain text-mining rules and then the importance weights of these words were calculated by using the inverse document frequency method. The results shows that some words such as "soil", "treatment, and "emission" have become outdated in recent years, despite their foremost importance. In addition, it has been observed that some concepts with relatively low levels of importance, such as "climate change" or "remote sensing", may appear as current issues.

Meanwhile, similar patterns were observed in the results obtained. Although Concepts 1, 3, and 5 have a positive trend, Concepts 2 and 4 are in a negative trend and can be

considered out of date. In addition, although Concepts 20, 22, and 22 have relatively low importance weights, it has been observed that they cover trending topics.

The results showed that unstructured texts can be converted into structured numerical data by analyzing academic publications with text mining, and thus meaningful data can be extracted with various statistical analyses. The findings of this study may help to direct new studies, to divide fund transfers into more current and trending areas, and even to test the topicality of studies at the evaluation stage in academic publication.

Funding: This research was funded by İskenderun Technical University—Scientific Research Projects Unit, grant number 2021YP01.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this study can be accessed from the Web of Science database. Available online: https://www.webofscience.com (accessed on 22 November 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Global Forest Watch. Available online: https://www.globalforestwatch.org/dashboards (accessed on 22 November 2022).
2. Web of Science. Available online: https://www.webofscience.com (accessed on 22 November 2022).
3. Ananiadou, S.; Rea, B.; Okazaki, N.; Procter, R.; Thomas, J. Supporting systematic reviews using text mining. *Soc. Sci. Comput. Rev.* **2009**, *27*, 509–523. [CrossRef]
4. Babić, D.; Kalić, M. Modeling the Selection of Airline Network Structure in a Competitive Environment. *J. Air Transp. Manag.* **2018**, *66*, 42–52. [CrossRef]
5. Jun, S.; Park, S.S.; Jang, D.S. Document Clustering Method Using Dimension Reduction and Support Vector Clustering to Overcome Sparseness. *Expert Syst. Appl.* **2014**, *41*, 3204–3212. [CrossRef]
6. Monali, P.; Sandip, K. A Concise Survey on Text Data Mining. *Int. J. Adv. Res. Comput. Commun. Eng.* **2014**, *3*, 8040–8043. [CrossRef]
7. Him, J.-G.; Ryu, K.-H.; Lee, S.H.; Cho, E.-A.; Lee, Y.J.; Ahn, J.H. Text Mining Approaches to Analyze Public Sentiment Changes Regarding COVID-19 Vaccines on social media in Korea. *Int. J. Environ. Res. Public Health* **2021**, *18*, 6549. [CrossRef] [PubMed]
8. Kitsios, F.; Kamariotou, M.; Karanikolas, P.; Grigoroudis, E. Digital Marketing Platforms and Customer Satisfaction: Identifying eWOM Using Big Data and Text Mining. *Appl. Sci.* **2021**, *11*, 8032. [CrossRef]
9. Atay, M.; Eroğlu, Y.; Ulusam Seçkiner, S. Investigation of breaking points in the airline industry with airline optimization studies through text mining before the covid-19 pandemic. *Transp. Res. Rec.* **2021**, *2675*, 301–313. [CrossRef]
10. Eroglu, Y.; Seçkiner, S.U. Trend Topic Analysis for Wind Energy Researches: A Data Mining Approach Using Text Mining. *J. Technol. Innov. Renew. Energy* **2016**, *5*, 44–58. [CrossRef]
11. Ertek, G.; Kailas, L. Analyzing a Decade of Wind Turbine Accident News with Topic Modeling. *Sustainability* **2021**, *13*, 12757. [CrossRef]
12. Mustaqim, T.; Umam, K.; Muslim, M.A. Twitter text mining for sentiment analysis on government's response to forest fires with vader lexicon polarity detection and k-nearest neighbor algorithm. *J. Phys. Conf. Ser.* **2020**, *1567*, 032024. [CrossRef]
13. Miner, G.; Elder, J.; Hill, T.; Nisbet, R.; Delen, D.; Fast, A. *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*, 1st ed.; Academic Press: Cambridge, MA, USA, 2012.
14. Delen, D.; Crossland, M.D. Seeding the survey and analysis of research literature with text mining. *Expert Syst. Appl.* **2008**, *34*, 1707–1720. [CrossRef]
15. Miller, T.W. *Data and Text Mining: A Business Applications Approach*; Pearson Prentice Hall: Upper Saddle River, NJ, USA, 2005.
16. Baker, K. *Singular Value Decomposition Tutorial*; The Ohio State University: Columbus, OH, USA, 2005; Volume 24.