Mapping the Research Landscape of Agricultural Sciences

Dmitry Devyatkin [a]

Junior Research Fellow, devyatkin@isa.ru

Elena Nechaeva [b]

Secretary of the Council for Implementation of the Federal Programme of Science and Technology Development in Agriculture for the years 2017–2025, valkmanalena@rambler.ru

Roman Suvorov [a]

Junior Research Fellow, rsuvorov@isa.ru

Ilya Tikhomirov [a]

Head of Laboratory "Intelligent Technologies and Systems", tih@isa.ru

[a] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, 9, 60-letiya Oktyabrya ave., Moscow 117312, Russian Federation

[b] Administration of the President of the Russian Federation, 4, Staraya sq., Moscow 103132, Russian Federation

Abstract

A research landscape is a high-level description of the current state of a certain scientific field and its dynamics. High-quality research landscapes are an important tools that allow for more effective research management. This paper presents a novel framework for the mapping of research. It relies on full-text mining and topic modeling to pool data from many sources without relying on any specific taxonomy of scientific fields and areas. The framework is especially useful for scientific fields that are poorly represented in scientometric databases, i.e., Scopus or Web of Science. The high-level algorithm consists of (1) full-text collection from reliable sources; (2) the automatic extraction of research fields using topic modeling; (3) semi-automatic linking to scientometric databases; and (4) a statistical analysis of metrics for the extracted scientific areas. Full-text mining is crucial due to (a) the poor representation of many Russian research areas in systems like Scopus or Web of Science; (b) the poor quality of Russian Science Index data; and (c) the differences between taxonomies used in different data sources. Major advantages of the proposed framework include its data-driven approach, its independence from scientific subjects' taxonomies, and its ability to integrate data from multiple heterogeneous data sources. Furthermore, this framework complements traditional approaches to research mapping using scientometric software like Scopus or Web of Science rather than replacing them. We experimentally evaluated the framework using agricultural science as an example, but the framework is not limited to any particular domain. As a result, we created the first research landscape covering young researchers in agricultural science. Topic modeling yielded six major scientific areas within the field of agriculture. We found that statistically significant differences between these areas exist. This means that a differentiated approach to research management is critical. Further research on this subject includes the application of the framework to other scientific fields and the integration of other collections of research and technical documentation (especially patents).

Keywords: text mining; topic modelling; science mapping; scientific landscape; agricultural science; publication activity; scientometrics; young researchers; Russian Science Index.

Article type: research paper

Citation: Devyatkin D., Nechaeva E., Suvorov R., Tikhomirov I. (2018) Mapping the Research Landscape of Agricultural Sciences. *Foresight and STI Governance*, vol. 12, no 1, pp. 57–66. DOI: 10.17323/2500-2597.2018.1.57.66

The research landscape reveals the structure of and major trends in knowledge creation [*Oldham et al.,* 2012; *Christofilopoulos, Mantzanakis,* 2016]. The process of constructing it is often referred to as mapping studies. These terms are similar to the generally accepted concepts of "patent landscape" and "patent mapping"[1]; the only difference is, the former are applied to academic publications as opposed to patents.

Information and analytical systems are usually used to map patent and scientific landscapes, such as Google Patents, PatSearch, Exactus Patent, Scopus, Web of Science, etc. However, they require having a representative base of documents, which, for various reasons, is not always possible, for example, Russian agricultural studies are poorly reflected in international scientometric databases. According to InCites, the number of agriculture-related papers indexed in the Web of Science Core Collection (WoS) and published between 2012–2016 is no more than 2,000 (the results differ depending on the selected classification). The most popular and most rapidly growing area is soil science or pedology. Also, agricultural sciences are represented 90% less often than medicine- and health-related sciences. A random check revealed that very few Russian doctorate holders specializing in this area have even a single publication indexed in the WoS or Scopus. Therefore, these databases are unsuitable for mapping the landscape of Russian agricultural sciences.

An alternative to international scientometric databases is the Russian Science Citation Index (RSCI) database, which more accurately reflects the structure of Russian science [*Zibareva et al.,* 2015]. However, the RSCI in its current state cannot be used as the sole data source for mapping a scientific landscape for several reasons [*Eremenko,* 2014; *Fradkov,* 2015]. Inaccurate links between publications and authors, difficulties surrounding the correction of such shortcomings, and the inclusion of all scientific publications in the RSCI regardless of their quality (which potentially allows one to "beef up" one's figures) present serious problems. For example, not so long ago, more than 300 low-quality publications were excluded from the RSCI, but the quality of the remaining ones has not been assessed [*Ekonomov,* 2017]. To prevent the artificial "inflation" of the scientometric figures, the RSCI adopted more stable versions of popular scientometric indicators: h-index without self-citation, core RSCI h-index, and the Herfindahl Index for journals. However, the quality of the primary data (and in particular the links between authors and publications) still leaves much to be desired. A number of other issues were revealed in the course of our study, which will be described below. The aforementioned limitations may be explained by low publication activity, insufficiently developed citation culture in agricultural sciences, and the RSCI's own flaws. Other researchers have also come to similar conclusions [*Sidorova,* 2016].

To deal with at least some of these problems, we suggest using additional data sources, in particular data about defended theses available at the State Supreme Certification Commission (VAK). According to Russian legislation, researchers' qualifications are assessed in the scope of the national attestation system, which awards Candidate of Sciences and Doctor of Sciences academic degrees in line with the nomenclature of the research specializations. It can be argued that a researcher who defended a thesis in the relevant scientific domain does have certain competences in the field — which should improve the quality of the research landscape. Also, the VAK database includes data about researchers' ages, which can be taken into account when conducting an analysis, thus ensuring a focus on the activities of young researchers.

Another common problem with mapping research landscapes is different classifications, which are supposed to help with the analysis by structuring subject areas but, in effect, only hinder it. The Russian State Classification of S&T Information (RCSTI), the Universal Decimal Classification (UDC), and the VAK nomenclature of research specializations could be used to analyze agricultural sciences. However, these classifications are not coordinated with one another and are very different. They also evolve: with time some codes are excluded and new ones are introduced. Plus, the subject area under consideration does not always fit the taxonomy suggested by the classifications — not by far. Accordingly, assigning codes to specific studies can produce very questionable results, and not infrequently, incorrect ones. The above issues can be dealt with by applying computerized text analysis, in particular clustering and thematic modeling techniques [*Shvets et al.,* 2015]. These approaches allow one to structure subject areas without using any existing classifications. When such tools are applied, the initial set is divided into groups on the basis of specified criteria.

In this study we have grouped the abstracts of Russian researchers' candidate and doctor theses on the basis of a full-text analysis. The resulting groups are seen as *research areas*.

The landscape mapping process comprised of three main stages:

- drafting the initial list of researchers and indicators of their activities;
- grouping the researchers by research area;
- comparing the groups and analyzing the indicators' dynamics.

The suggested methodology was tested for young researchers specializing in agricultural sciences, wich is particularly relevant for knowledge areas poorly represented in Scopus and the WoS. The computerized

[1] Order of the Russian Federal Service for Intellectual Property No. 8 from 01.01.2017. Available at: http://www.consultant.ru/document/cons_doc_LAW_212062/, accessed 15.12.2017.

tools for analyzing native-language texts that the methodology is based upon allow one to aggregate data about researchers' publications and thesis-writing activities from various sources, regardless of how structured it happens to be.

Initial Data

The abstracts of candidate and doctor theses, the Russian Ministry of Education and Science's information about researchers awarded academic degrees in agricultural sciences (full name, organisation, date of birth (VAK data)), and the RSCI scientometric data (http://elibrary.ru) provided the informational basis for this study.

The initial sample was comprised of researchers under 40 years of age (at the time of the study, i.e., in 2016), who have defended a thesis in agricultural sciences between 2008-2015, inclusively. The limited timeframe for the theses defense was due to a lack of the VAK data from earlier periods. Altogether, the sample included 2,572 young researchers.

The data provided by the VAK included four top-level research specialization codes: 06 (agriculture), 05 (engineering sciences), 03 (biological sciences), and 25 (Earth sciences); 10 second-level codes, and 32 third-level ones (full research specialization identifiers). Interestingly, some of the codes belonged to the old version of the VAK classification, while others belong to the new one, so applying them to map a research landscape without additional processing was very problematic.

A set of full-text abstracts of candidate and doctor theses collected on the website of the Russian State Library was used to find out the structure of the scientific domain under consideration. The abstracts were linked to the RSCI and VAK data using the researchers' full names and the names of the organizations where they conducted their dissertation studies.

The researchers' scientometric indicators were collected automatically. A full-name search of the eLIBRARY author index was conducted, followed by searching for their organisations (relevant fields of the authors' RSCI pages and the VAK database were compared). Year of birth and first publication date were also taken into account in order to exclude profiles with incorrectly linked publications. If all of the above conditions were met, the software module was launched to import the relevant scientometric indicators into the database containing the initial data for subsequent landscape mapping. Working with the RSCI information system revealed a number of its flaws:

- a small share of updated author profiles (only 56%);
- lack of data about researchers' age (which does not allow one to analyze young researchers on their own);
- lack of sufficiently powerful analytical tools or data uploading mechanisms for subsequent analysis using other tools;
- low informative value of core RSCI publications' h-index (for 75% of the analyzed researchers its value did not exceed 1);
- a single thematic classification, which does not allow one to categorize the research areas and analyze them from various perspectives. To compare, the aforementioned WoS does offer an opportunity to use different classifications to analyze a particular area's structure.

The chi-square test revealed no statistically significant differences in automatically generated research areas' distribution in the general population (2,572 researchers), and in the sample (56% of the authors with successfully updated RSCI profiles). This is grounds to believe that the sample was representative of the initial list of researchers drafted on the basis of VAK data. All subsequent operations with scientometric indicators were performed only for the 56% of researchers with successfully linked RSCI profiles.

The initial data set was presented in the form of a table with the following columns:

- full name;
- year of birth;
- academic degree;
- year of defending thesis;
- specialization code;
- organization;
- thesis abstract identifier;
- number of publications;
- h-index;
- h-index without self-citation;
- h-index for core RSCI publications;
- total number of citations;
- average weighted impact factor of journals where the papers were published;
- average weighted impact factor of journals where the papers were cited;
- research areas.

Each row of the table corresponds to a particular researcher. Data in the field "Research areas" was entered in line with the structure revealed by thematic modeling of the set of full-text abstracts.

Establishing the Thematic Structure of the Subject Area

In order to identify agricultural sciences' research areas on the basis of the set of full-text abstracts of young researchers' theses obtained from the VAK database, a probabilistic thematic model of the latent Dirichlet allocation (LDA) was built [*Blei et al.,* 2003]. This methodology has repeatedly allowed one to correctly interpret the results of grouping similar texts, which was reflected in a number of Russian and international studies [*Garousi, Mäntylä,* 2016; etc.]. It provides an opportunity to obtain a generalized description of the thematic structure of the text array being analyzed. The Gensim library software version of LDA was used during our experiment [*Rehurek, Sojka,* 2010].

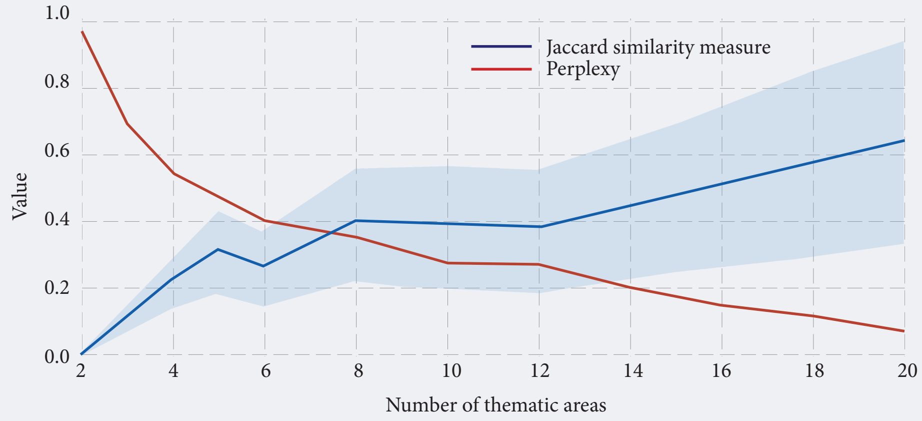
The set of attributes was selected using the "bag-of-words" approach. Since the abstracts contained no author-specific keywords or phrases, the only way to determine the studied texts' vocabulary was by analyzing their full texts. A linguistic analyzer was applied to extract relevant words and phrases from the full texts of the abstracts and was capable of conducting morphologic, syntactic, and semantic analysis, including semantic role labeling, for texts in Russian and English. The morphology of Russian-language texts was processed using the ATP library techniques [*Sokirko,* 2001]; the syntax — with the help of the MaltParser software [*Nivre et al.,* 2007] pre-trained using a marked-up SynTagRus array [*Nivre et al.,* 2008]. Similar procedures for English-language publications were performed using Freeling library tools [*Padró, Stanilovsky,* 2012]. The relational-situational analysis technique developed by the Systems Analysis Institute of the Federal Research Centre for Informatics and Management of the RAS [*Osipov et al.,* 2013], which provided a basis for processing semantics. An assessment of results of a marked-up array analysis are presented in [*Shelmanov, Smirnov,* 2014]. The high efficiency of the applied analyzer is indirectly confirmed by the results of its previous practical applications, including its ability to reveal text borrowing (PAN CLEF-2014) [*Zubarev, Sochenkov,* 2014] and to conduct a full-text search (ROMIP-2008) [*Smirnov et al.,* 2008]. The tool allows one to identify specific words and name groups combining syntactically linked words, in which a noun plays the leading role [*Suvorov, Sochenkov,* 2015]. The resulting phrases are normalized by bringing the main word to the normal form, after which the form of the dependent word is chosen from all possible variants whose morphological characteristics match those of the main word.

Perplexy was chosen as the main criterion to assess the thematic model's quality [*Hofmann,* 1999]. It describes the model's ability to restore the initial probabilistic distribution of the analyzed document set's vocabulary (words and phrases): the lower this indicator's value, the better the model describes the data. The most significant parameter for building the model is the number of thematic areas (or, in other words, classes). The higher this value, the more flexible the model becomes, so it will describe the initial data increasingly better (therefore, the perplexy value decreases). However, if we set an excessively high value, thematic areas would significantly overlap and the breakdown would be hard to interpret. Accordingly, the issue of setting the right number of thematic areas arises. In this study, the Jaccard measure was applied to assess thematic areas' similarity [*Manning et al.,* 2008] by using lists comprising 30 words and phrases most typical for each area. The best number would be the one providing a sufficiently low perplexy figure, combined with a not too high degree of similarity.

We shall use the term *research areas* (or just *areas*) to refer to the results of thematic modeling conducted in line with the above scheme. Following this procedure, some of the abstracts could be included in several areas at the same time, which provides an opportunity to identify and analyze groups of interdisciplinary studies.

Figure 1 presents graphs of the thematic model's normalized perplexy, the normalized average value of the Jaccard similarity measure for the areas, and the dispersion of these values. When the number of areas

Figure 1. **Perplexy and Jaccard Similarity**



Source: composed by the authors.

Table 1. **Research Areas Automatically Generated by Processing Full-Text Abstracts Published in 2008–2015**

Research area	Keywords and phrases (automatically generated)
Meat and dairy animal farming	cow, animal, milk, cattle, live weight, lactation, bull calf, milk yield, genotype, milk production, monthly index, live, cow group, female, cross-breed, dairy, line, goal, contemporary, Indian ink, monthly age, mare, steppe, Holstein-Friesian breed, blood, herd, ram, heifer, first calf heifer, birth, youngsters, animal group, meat production, animal, head, contemporary, meat
Plant selection	hybrid, line, score, fruit, standard, sort sample, population, disease, stock, coastal, early, combination, sprout, grain weight, new variety, ear, genotype, decade, fruity, selection, variability, economic, blossom, apple tree, seed weight, planting, potato, cross-breeding, shoot, create, collection, original material, damage, diameter, soil, maturing, nursery, lodging, garden
Forage	ration, experimental group, control group, mixed fodder, live weight, additive, feeding, animal, broiler, protein, digestibility, live, organism, preparation, broiler chicken, bird, blood, fodder calcium, chicken, piglet, preservation, egg, raw protein, feeding, head, analogue, metabolism energy, scientific and economic experiment, pig, cubs, goal, inclusion, piggy, chicken, cellulose, meat, duration of experiment, substance, digestibility, methionine
Silvics	plantation, tree, forest, pine, wood, earth, cutting, oak, diameter, forest stand, stock, incline, class, belt, pinery, tier, care, fir grove, timber, cover, agro-landscape, burbot, distribution, fir, undergrowth, crown, forest shelter belt, forest type, cone, timberland, horizon, model, fescue, herbage, northern, tillage, common pine
Pedology	irrigation, quantity, watering, edge, preparation, solution, heavy metal, loss, fraction, July, rice, irrigation regime, root rot, pest, plant louse, acid, incentive, Kuban, amaranth, mine, environmentally, larva, speed, September, biologically, water regime, horizon
Field crops	winter wheat, mineral fertilizer, barley, spring wheat, kg/ha, crop rotation, cwt/ha, seeding rate, sowing time, green mass, grass, rye, herbicide, fallow, predecessor, fertilizer application, corn, mix, winter rye, seed processing, pea, manure, oats, tillage, lucerne, preparation, perennial grass, nitrogen fertilizer, black earth, ploughing, soil layer, soya, fertilizer dose, grain quality, clover, germination, l/ha, crop capacity, dressing, seeding technique

Source: composed by the authors.

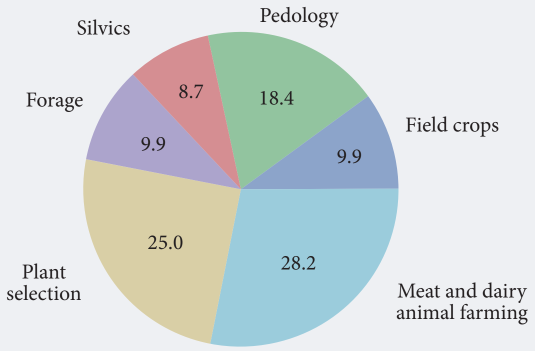
grows from 6 to 8, the dispersion of the Jaccard similarity measure significantly increases: by more than 50%. Pairs of areas emerge with a high share of common vocabulary, while perplexy drops insignificantly. This and the interpretation of the thematic breakdown clearly suggest that the best number of areas would be six. It ensures that the resulting thematic breakdown can be properly interpreted and provides an acceptable balance of perplexy and Jaccard similarity values.

Names of the areas, together with corresponding typical keywords and phrases, are presented in Table 1. They were formulated by the authors on the basis of the lists of keywords and phrases automatically generated during the testing of the thematic LDA model.

Figure 2 shows the shares of theses defended in each of the automatically generated research areas. In comparison, Figure 3 presents the distribution of this indicator by 15 agricultural research specialization nomenclature codes with the highest activity.

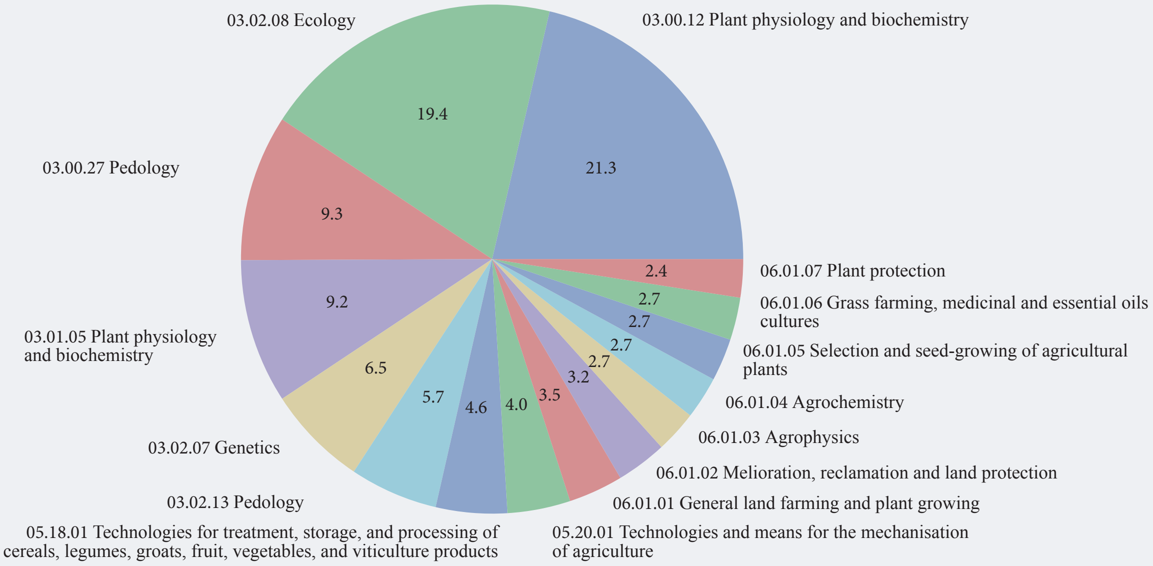
The distribution presented in Figure 3 is rather skewed (few popular specialization areas and a long "tail" of those with few defended theses). It would be hard to conduct a comparative analysis of research areas on this basis for several reasons. Since the list of research specializations simultaneously contains old and new VAK nomenclature codes, comparing them turns into a formidable challenge with no definitive solution. Also, due to significantly different sample sizes for each specialization, estimates of scientometric indicators' distribution parameters have different confidence intervals (and for less popular specializations, they become increasingly larger). For particularly rare research areas, estimating such

Figure 2. **Distribution of Defended Theses by Research Areas (%)**



Source: composed by the authors.

Figure 3. **Distribution of Defended Theses by Top 15 Agricultural Research Specialisation Nomenclature Codes** (%)



Source: composed by the authors.

parameters becomes totally futile, which makes it impossible to compare them with better-represented research areas.

Comparing Figures 2 and 3 reveals that the distribution of defended theses by automatically generated research areas is less skewed than the one by research specialization codes. This simplifies further thematic analysis and its validity for structuring research areas and assessing publication activity indicators can be considered confirmed. An important difference between research areas and research specialization codes is that the former overlap: any researcher may belong to several research areas at the same time. Accordingly, the study takes interdisciplinarity into account, which is a major characteristic of present-day research.

Figure 4 shows the annual numbers of defended agricultural sciences-related theses between 2008-2015 inclusively. The columns' colored sections indicate the shares of theses defended in each specific research area during the relevant year. We can see that the emphasis is shifting from field crops and pedology towards plant selection.

Analysis of Young Researchers' Scientometric Indicators

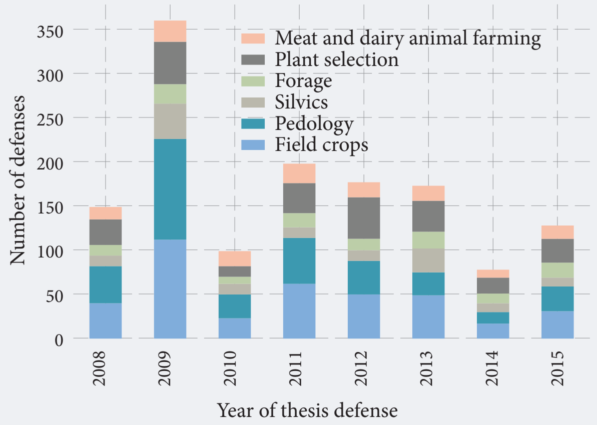
The scientometric indicators were analyzed using simple histograms — calculations of the density distributions with Gaussian smoothing [*Scott,* 1992]. The comparison graphs are presented as violin plots [*Hintze, Nelson,* 1998]. The Python 2.7 programming language and the Jupyter interactive development environment were applied for data processing purposes. The graphs were built using the Matplotlib [*Hunter,* 2007] and Seaborn (https://seaborn.pydata.org/) libraries.

As noted, less than 56% of young Russian researchers specializing in agricultural sciences have updated RSCI profiles. This means that they pay very little attention to monitoring their publication activity. After the eLIBRARY profiles were linked with the VAK data, records containing all known parameters suitable for further analysis remained for only 1,419 researchers. It turned out that only 107 researchers had an h-index value of 5 or above (out of the more than 2,500 researchers on the initial list). As to the h-index without self-citation, only 78 young researchers have it at 5 or above.

Figure 5 shows the correlation between the researchers' age and number of publications. Each dot corresponds to a specific researcher (their age is shown on the horizontal axis, and the number of publications at the time of the study — on the vertical one). The linear trend (correlation) is shown as a straight line. The oval lines describe density of the joint "age and number of publications" distribution.

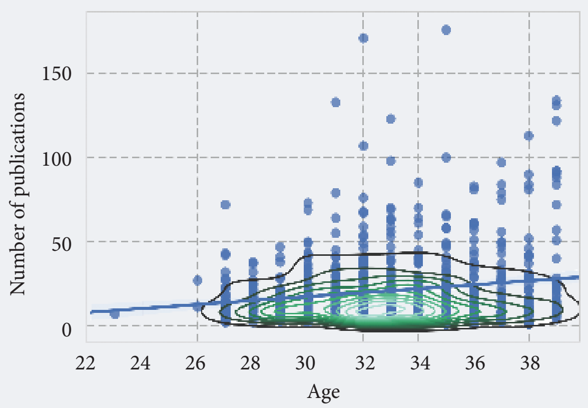
From a demographic point of view, the graph in Figure 5 shows that researchers specializing in agricultural sciences are most productive at the ages of 32–34. Typically, they are young, actively working, and publishing candidates of sciences. Another spike in activity occurs at the age of 39–40; presumably this is the most productive period of young doctors of sciences or candidates getting ready to defend their doctoral theses. The 30-year threshold is preceded by a smooth growth of publication activity, which stabilizes during the next five years.

Figure 4. **Dissertation Activity Growth by Research Area**



Source: composed by the authors.

Figure 5. **Linear Correlation and Density of Joint Distribution of Researchers' Age and Number of Publications in the RSCI**



Source: composed by the authors.

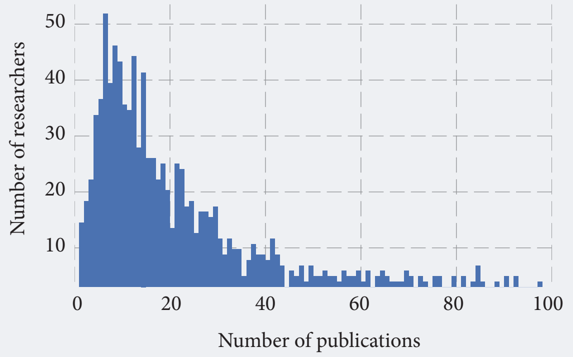
Figure 6 presents a histogram of distribution of the number of publications (limited to 100). Most of the researchers have between 8 and 25 publications to their credit; the largest number is 176.

Figure 7 shows histogram of h-index distribution without self-citation. We can see that most of the researchers have an h-index between 1 and 3. The highest value in the sample is 22.

Figures 8–11 present violin plots showing the empirical distribution of the various scientometric indicators between automatically generated research areas. Each figure represents a graph of core probability density estimation (symmetrical in relation to the vertical axis), with Gaussian smoothing. The wider the figure, the higher the share of researchers with the corresponding index value. Dotted lines inside the figures mark the 25th, 50th, and 75th percentiles. To assess the statistical significance of the differences in distributions by specific research areas, the Mann–Whitney U-test was applied [*Mann, Whitney,* 1947]. The objective was to make sure that in some areas the average indicator value was higher than in others, so the application of that specific criterion seems to be valid. The criterion value and the corresponding p-value were calculated for all possible research area pairs (15 altogether). The statistical significance threshold (critical p-value) was set at 0.05.

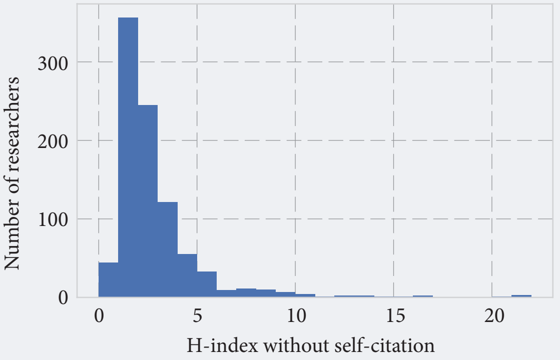
Figure 8 presents distribution of the h-index (without self-citation) by automatically generated research areas. The figures' width is proportional to the number of researchers with corresponding h-index values. Researchers whose h-index is below 5 are not shown on the graph. They were cut from the sample to show the difference between the levels of leading researchers' citations in each research area. The graph shows that world-class researchers (h-index of 10 or more) are represented in various research areas to differing degrees. The application of the Mann–Whitney criterion[2] established that if the significance threshold is

Figure 6. **Histogram of Number of Publications Distribution by Researcher**



Source: composed by the authors.

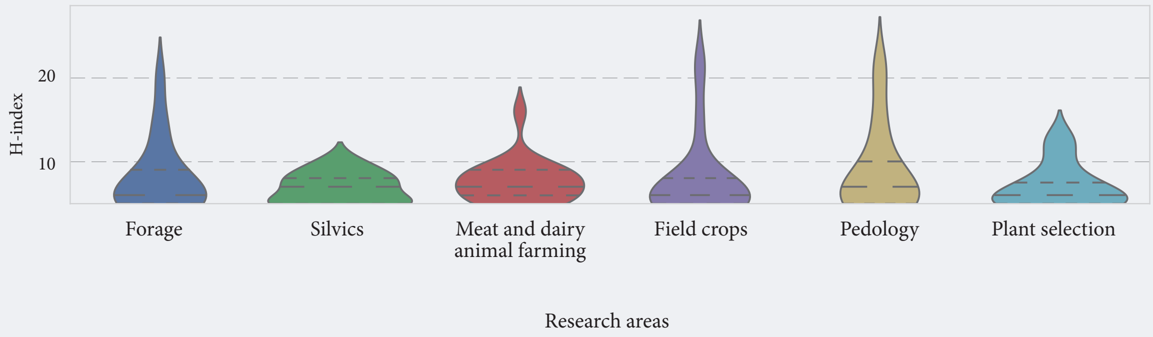
Figure 7. **Histogram of H-index Distribution (Without Self-Citation)**



Source: composed by the authors.

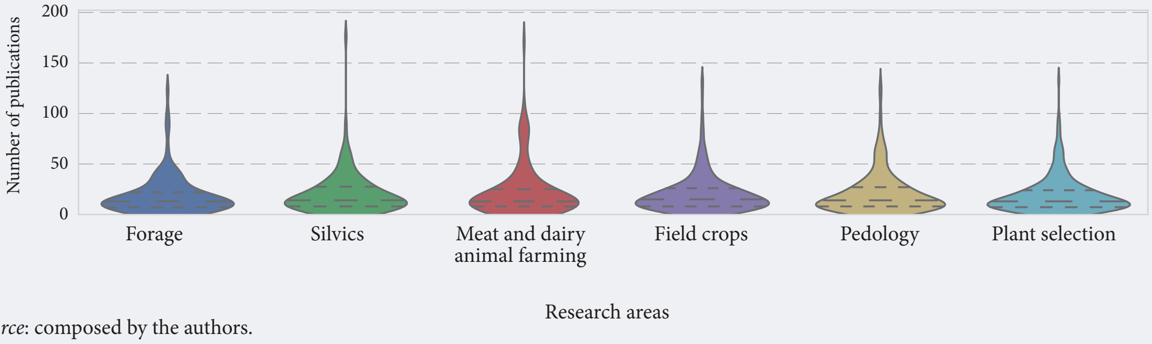
[2] The Mann–Whitney criterion is calculated based on the basis of all h-index values (including those under 5).

Figure 8. **H-index Distribution (Without Self-Citation)**



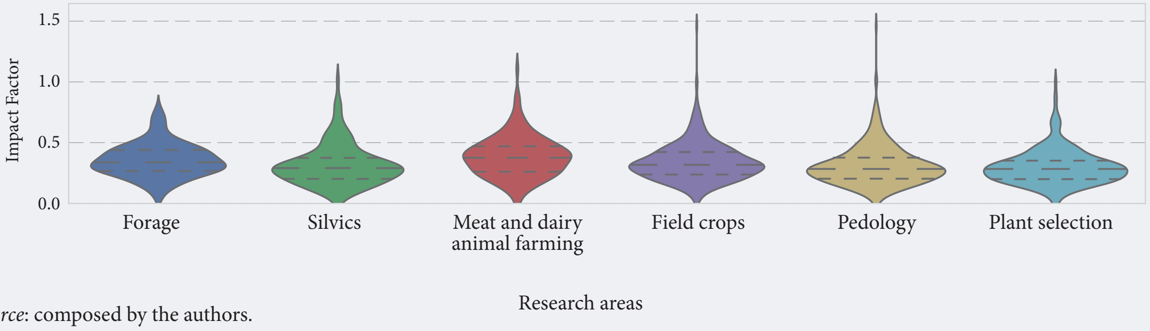
Source: composed by the authors.

Figure 9. **Number of Publications Distribution**



Source: composed by the authors.

Figure 10. **Distribution of Average Weighted Impact Factors of the Journals Where the Researchers' Papers Were Published**



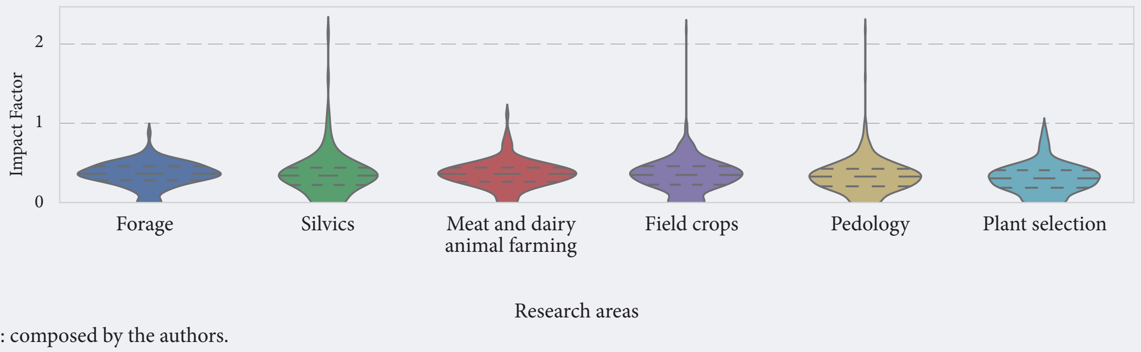
Source: composed by the authors.

set at 0.05, in 40% of all cases (six research area pairs) the average h-index in one area was higher than in the other one. The most significant divergence was noted for the following pairs: "silvics – meat and dairy animal farming" (p=0.007), "meat and dairy animal farming – plant selection" (p=0.0003), and "field crops – plant selection" (p=0.0006). The least significant difference was found for the pair "forage – meat and dairy animal farming" (p=0.44).

Figure 9 shows distribution of researchers' publications in each area. We can see that all researchers' publication activity remains at a similar level. No statistically significant variations between the average numbers of publications in each area were discovered (average p=0.28).

Figures 10 and 11 show the impact factors of the journals that publish the researchers' papers and that of the publications which cite these papers. We can see from the graphs that the distributions in the various research areas are quite different. Statistically significant variation between impact factors of the journals where the papers were published and the journals where they were cited have been found for ten and seven research area pairs, respectively (67% and 47% of the pairs). The most significant variation of the average indicator values were found for the following pairs: "meat and dairy animal farming — plant selection" (p=0.0005), "forage — plant selection" (p=0.0004), "field crops — plant selection" (0.0008).

Figure 11. **Distribution of the Average Weighted Impact Factors of Journals That Cited the Researchers' Papers**



Source: composed by the authors.

The least significant difference was noted for the combinations "forage — meat and dairy animal farming" (p=0.37) and "silvics — field crops" (p=0.32).

It would be interesting to consider the differences between the impact factors of the journals that have published the papers and the ones that have cited them in the scope of the same research area. The graphs show that, for example, the journals that cite papers on silvics and pedology (unlike other research areas), have a much higher rating than the journals which originally published the papers. This was confirmed by the Mann-Whitney U-test, with p=0.03 and p=0.02, respectively.

The above gives firm ground to conclude that certain differences exist between particular research areas of agricultural sciences in terms of scientometric indicators and citation alike. This is suggested by the results of the statistical Mann-Whitney U-test, which, in turn, confirm the need to use different approaches when analyzing the productivity of researchers specializing in different research areas in the scope of the same scientific discipline.

Conclusion

The paper proposes and tests a new research landscape mapping methodology. Its originality lies in the combination of full-text analytics and traditional statistical processing of scientometric data to improve the reliability, sustainability, and interpretability of research landscapes. The proposed approach corrects the flaws of manually constructed taxonomies which tend to lean towards an excessive level of detail and offer little opportunity for one to compare research areas with one another. The methodology is especially relevant for disciplines poorly represented in international scientometric databases such as Scopus and the WoS. Note that the proposed technique for combining various data sources does have alternatives, but in the authors' opinion, it does help one solve the various problems described in the paper.

The suggested toolset does not replace traditional scientometric tools but supplements them, allowing one to obtain a more holistic and more easily interpretable picture in order to accomplish specific objectives in the area of analyzing and assessing scientific development. Its advantages include an opportunity to identify emerging prospective interdisciplinary research areas (a "data-based" approach). Accordingly, it can be applied, for example, to address global challenges mentioned in the S&T Development Strategy of the Russian Federation[3].

The experimental testing of the methodology for the first time allowed us to map the research landscape of the agricultural sciences, using young researchers aged under 40 as a sample. Six major research areas were identified: forage, silvics, meat and dairy animal farming, field crops, pedology, and plant selection. The highest values of bibliometric indicators were noted in pedology. The significant divergence of the abovementioned areas' scientometric indicators suggest a need to use individual approaches when assessing relevant research results.

In the authors' opinion, the prospects for the further development of this methodology include mapping other research landscapes, designing algorithms for identifying promising research areas, and building area-specific rankings of researchers and R&D organizations. The scope for the application of text analysis techniques to compare S&T documents from various sources must be explored (such as databases of academic papers, patents, etc.), which cannot be automatically linked to each other due to the lack of a comprehensive classification. Such a comparison could provide a basis for the integrated analysis of specific S&T areas and for mapping more complex research landscapes.

The study was sponsored by the RFBR, grant № 14-29-05008.

[3] Approved by the RF Presidential Decree No. 642 of 01.12.2016 . Available at: http://publication.pravo.gov.ru/Document/View/0001201612010007, accessed 14.11.2017.

References

Blei D.M., Ng A.Y., Jordan M.I. (2003) Latent Dirichlet Allocation. *Journal of Machine Learning Research*, vol. 3, pp. 993–1022.

Christofilopoulos E., Mantzanakis S. (2016) China 2025: Research & Innovation Landscape. *Foresight and STI Governance*, vol. 10, no 3, pp. 7–16. DOI: 10.17323/1995-459X.2016.3.7.16.

Ekonomov I. (2017) Musornaya nauka [Junk science]. *TrV-Nauka* [TrV-Science]. Available at: http://trv-science.ru/2017/04/25/musornaya-nauka/, accessed 15.05.2017 (in Russian).

Eremenko G. (2014) Vo vsem vinovat RINTs? [Is everything to blame Russian Science Citation Index?]. *Troitskii variant*, no 163, p. 7 (in Russian).

Fradkov A. (2015) RINTs prodolzhaet vrat' [Russian Science Citation Index]. *TrV-Nauka* [TrV-Science]. Available at: http://trv-science.ru/2015/09/08/risc-prodolzhaet-vrat/, accessed 15.05.2017 (in Russian).

Garousi V., Mäntylä M.V. (2016) Citations, research topics and active countries in software engineering: A bibliometrics study. *Computer Science Review*, vol. 19, pp. 56–77.

Hintze J.L., Nelson R.D. (1998) Violin plots: A box plot-density trace synergism. *The American Statistician*, vol. 52, no 2, pp. 181–184.

Hofmann T. (1999) Probabilistic latent semantic analysis. *Proceedings of the XV Conference on Uncertainty in Artificial Intelligence*, Berlington, MA: Morgan Kaufmann Publishers, Inc., pp. 289–296.

Hunter J. D. (2007) Matplotlib: A 2D graphics environment. *Computing in Science and Engineering*, vol. 9, no 3, pp. 90–95.

Mann H.B., Whitney D.R. (1947) On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, vol. 18, no 1, pp. 50–60.

Manning C.D., Prabhakar R. Schütze H. (2008) *Introduction to information retrieval*, Cambridge: Cambridge University Press.

Nivre J. H., Hall J., Nilsson J., Chanev A., Eryiğit G., Kübler S., Marinov S., Marsi E. (2007) MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, vol. 13, no 2, pp. 95–135.

Nivre J., Boguslavsky I.M., Iomdin L.L. (2008) Parsing the SynTagRus treebank of Russian. *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*, Stroudsburg, PA: Association for Computational Linguistics (ACL), pp. 641–648.

Oldham P., Hall S., Burton G. (2012) Synthetic biology: Mapping the scientific landscape. *PLoS One*, vol. 7, no 4, p. e34368. Available at: https://www.ncbi.nlm.nih.gov/pubmed/22539946, accessed 24.05.2017.

Osipov G, Smirnov I., Tikhomirov I., Shelmanov A. (2013) Relational-situational method for intelligent search and analysis of scientific publications. *Proceedings of the Integrating IR Technologies for Professional Search Workshop, Moscow, Russian Federation, March 24, 2013* (eds. M. Lupu, M. Salampasis, N. Fuhr, A. Hanbury, B. Larsen, H. Strindberg), pp. 57–64. Available at: http://ceur-ws.org/Vol-968/irps_10.pdf, accessed 24.05.2017.

Padró L., Stanilovsky E. (2012) Freeling 3.0: Towards wider multilinguality. *Proceedings of the International Conference on Language Resources and Evaluation "LREC2012", Istanbul, 2012*, pp. 2473–2479. Available at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf, accessed 24.05.2017.

Rehurek R., Sojka P. (2010) Software framework for topic modelling with large corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, May 22*, pp. 51–56. Available at: http://www.lrec-conf.org/proceedings/lrec2010/workshops/W10.pdf, accessed 24.05.2017.

Scott D.W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*, New York; Chicester: John Wiley & Sons.

Shelmanov A.O., Smirnov I.V. (2014) Methods for semantic role labeling of Russian texts. *Computational Linguistics and Intellectual Technologies. Proceedings of International Conference Dialog*, vol. 13, no 20, pp. 607–620.

Shvets A., Devyatkin D., Sochenkov I., Tikhomirov I., Popov K., Yarygin K. (2015) Detection of Current Research Directions Based on Full-Text Clustering. *Proceedings of 2015 Science and Information Conference (SAI 2015), 28-30 July 2015, London, United Kingdom*, Piscataway, NJ: Institute of Electrical and Electronics Engineers (IEEE), pp. 483–488.

Sidorova V. V. (2016) Ispol'zovanie RINTs dlya otsenki nauchnoi deyatel'nosti gumanitariev [The use of Russian Science Citation Index for Evaluation of Humanities Research]. *Sibirskie istoricheskie issledovaniya* [Siberian Historical Research], no 1, pp. 27–39 (in Russian).

Smirnov I.V., Sochenkov I.V., Murav'ev V.V., Tikhomirov I. A. (2008) Rezul'taty i perspektivy poiskovogo algoritma Exactus [Results and prospects of the search algorithm Exactus]. *Trudy rossiiskogo seminara po otsenke metodov informatsionnogo poiska (ROMIP) 2007–2008* [Proceedings of the Russian seminar on the evaluation of information retrieval methods (ROMIP) 2007-2008], Saint Petersburg: Center for Strategic Research, pp. 66–76 (in Russian).

Sokirko A. (2001) *A short description of Dialing Project*. Available at: http://www.aot.ru/docs/sokirko/sokirko-candid-eng.html, accessed 15.05.2017.

Suvorov R.E., Sochenkov I.V. (2015) Establishing the similarity of scientific and technical documents based on thematic significance. *Scientific and Technical Information Processing*, vol. 42, pp. 321–327.

Zibareva I., Soloshenko N. (2015) Tematicheskaya struktura rossiiskogo segmenta nauchnykh zhurnalov v global'nykh i natsional'nykh informatsionnykh resursakh [Thematic structure of the Russian segment of scientific journals in global and national information resources] // Materialy Tret'ei mezhdunarodnoi konferentsii NEIKON 'Elektronnye nauchnye i obrazovatel'nye resursy: Sozdanie, prodvizhenie i ispol'zovanie' [Proceedings of the Third International Conference NEICON 'Electronic scientific and educational resources: Creation, promotion and use'], Moscow: NEICON, pp. 255–259 (in Russian).

Zubarev D., Sochenkov I. (2014) Using sentence similarity measure for plagiarism source retrieval. *Working Notes for CLEF 2014 Conference*, pp. 1027–1034. Available at: https://pdfs.semanticscholar.org/4556/08d685695c1a7f05ffd8257fae79e1f64593.pdf, accessed 15.05.2017.