RESEARCH ARTICLE
https://doi.org/10.17059/ekon.reg.2022-2-3
UDC 330.47; 332.1
JEL C38; D83; P25

Ivan V. Savin [a)], *Nikita S. Teplyakov* [б)]
[a, b)] Ural Federal University, Ekaterinburg, Russian Federation
[a)] Universitat Autonoma de Barcelona, Barcelona, Spain
[a)] https://orcid.org/0000-0002-9469-0510
[b)] https://orcid.org/0000-0003-2522-8207; e-mail: nekit_teplykov@mail.ru

Using Computational Linguistics to Analyse Main Research Directions in Economy of Regions[1]

Over the past decades, the process of knowledge generation has accelerated, producing a lot of scientific publications, which makes reviewing even a relatively narrow subject area very demanding, if not impossible. However, recent text data mining tools can assist researchers in conducting such analysis in an objective and time-efficient way. We conduct such a literature review on 1307 articles published in the journal Economy of Regions from 2010 to 2021 using advanced topic modelling techniques. This analysis aims to describe the main research areas in the journal over time, the dynamics of their popularity and the relationship with key quantitative indicators. We identified 22 topics ranging from "Agriculture" and "Economic Geography" to "Fiscal Policy" and "Entrepreneurship". We estimate how popularity of these topics was changing over time and find topics that gained the most popularity from 2010 to 2021 (+17.61 %, "Spatial Economics") or lost it (–14.58 %, "Economics of Innovation"). The topic of environmental economics collects the largest number of citations per article (3.64, on average), and the topics on monetary policy and poverty are the most popular among manuscripts in English, which is also true for articles written by authors with foreign affiliation. Papers with third-party funding are concentrated the most in "Spatial Economics" (around 11 %), and the least — in "Agriculture". Our results can help to understand the evolution in scope of research of Economy of Regions and serve researchers to find promising directions for future studies.

Keywords: topic modelling, machine learning, computational linguistics, text mining, literature review, academic journal, spatial economics, environmental economics, scientometrics, third-party funding

Acknowledgements

The article has been prepared with the support from the Russian Science Foundation, conducted as part of the research project № 19–18–00262 "Modelling a balanced technological and socio-economic development of the Russian regions".

For citation: Savin, I. V. & Teplyakov, N. S. (2022). Using Computational Linguistics to Analyse Main Research Directions in Economy of Regions. Ekonomika regiona [Economy of regions], 18(2), 338-352, https://doi.org/10.17059/ekon.reg.2022-2-3.

[1] © Savin, I. V., Teplyakov, N. S. Text. 2022.

ИССЛЕДОВАТЕЛЬСКАЯ СТАТЬЯ

И. В. Савин [а], Н. С. Тепляков [б]
[а, б] Уральский Федеральный университет им. первого Президента России Б. Н. Ельцина, г. Екатеринбург, Российская Федерация
[а] Автономный университет Барселоны, г. Барселона, Испания
[а] https://orcid.org/0000-0002-9469-0510
[б] https://orcid.org/0000-0003-2522-8207; e-mail: nekit_teplykov@mail.ru

Применение компьютерной лингвистики для анализа основных направлений исследований в журнале «Экономика региона»

За последние десятилетия процесс создания новых знаний значительно ускорился, что обусловило появление огромного количества научных публикаций. Это делает обзор даже относительно узкой предметной области крайне затруднительным. Тем не менее новейшие инструменты анализа текстовых данных могут помочь исследователям выполнить эту задачу объективно и с минимальными временными затратами. При помощи методов тематического моделирования мы проводим обзор литературы по 1307 статьям, опубликованным в журнале «Экономика региона» с 2010 г. по 2021 г. Данная работа нацелена на описание основных направлений исследований в журнале, динамики их популярности и взаимосвязи с ключевыми количественными показателями. В ходе анализа мы определили 22 основные темы исследований, варьирующихся от сельского хозяйства и экономической географии до фискальной политики и предпринимательства. Мы оценили, как со временем менялась распространенность этих тем, и определили тематики, которые либо набрали наибольшую популярность с 2010 г. по 2021 г. (+17,61 %, «Пространственная экономика»), либо потеряли ее (–14,58 %, «Экономика инноваций»). Статьи на тему экономики природопользования чаще цитируются (в среднем, 3,64 цитирований на 1 статью), а темы денежно-кредитной политики и бедности наиболее часто встречаются среди работ на английском языке, а также публикаций с иностранной аффилиацией. Работы, вышедшие при поддержке стороннего финансирования, наиболее сконцентрированы в теме «Пространственная экономика» (около 11 % статей), а наименее — в теме «Сельское хозяйство». Полученные результаты, демонстрирующие эволюцию исследований в журнале «Экономика региона», могут помочь авторам найти перспективные направления будущих работ.

Ключевые слова: тематическое моделирование, машинное обучение, компьютерная лингвистика, анализ текста, обзор литературы, научный журнал, пространственная экономика, экономика природопользования, наукометрия, стороннее финансирование

Благодарность

Исследование было выполнено в рамках научного проекта № 19-18-00262 «Моделирование сбалансированного технологического и социально-экономического развития российских регионов» при поддержке Российского научного фонда.

Для цитирования: Савин И. В., Тепляков Н. С. Применение компьютерной лингвистики для анализа основных направлений исследований в журнале «Экономика региона» // Экономика региона. 2022. Т. 18, вып. 2. С. 338-352. https://doi.org/10.17059/ekon.reg.2022-2-3.

Introduction

Over time, the amount of research studies is rapidly growing, making its review increasingly challenging when using conventional tools and human coders (Callaghan, Minx, Forster, 2020). Fortunately, over the last years, new methods from the intersection of machine learning and natural language processing (sometimes referred to as computational linguistics) have been developed. Using these methods to review publications and analyse trends in a scientific journal has become popular in the literature (Asmussen, Møller, 2019; Mo, Kontonatsios, Ananiadou, 2015; Savin, van den Bergh, 2021; Griffith, Steyvers, 2004; Lüdering, Winker, 2016; De Battisti, Ferrara, Salini, 2015; Ambrosino, et al., 2018; Maier et al., 2018).

To our knowledge, however, no study so far applied these methods to a Russian scientific journal[1]. We aim to fill this gap by analysing articles published in the journal Economy of Regions in the last twelve years.

Economy of Regions (ER, henceforth) is an international peer-reviewed journal founded in

[1] An exception may be Devitsyn and Savin (2020), but they analyse not a single journal but all publications by the authors affiliated with Surgut State University.

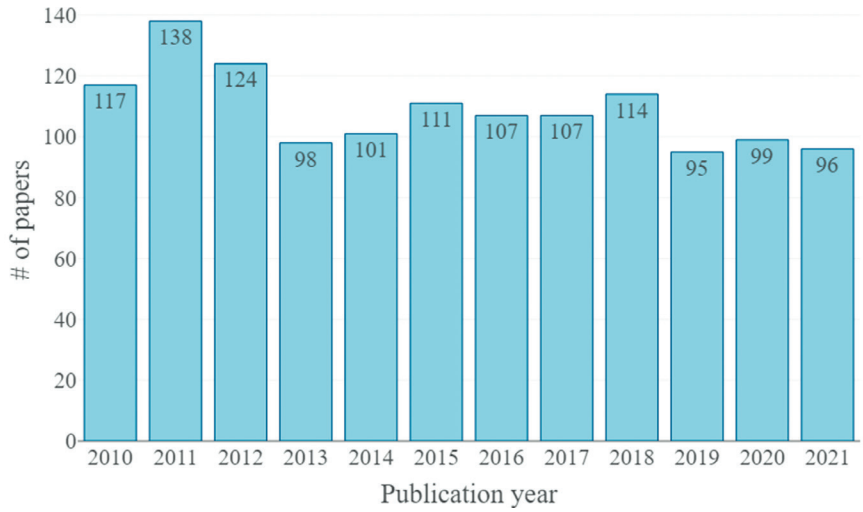

Fig. 1. *Number of articles published in ER over time in our sample*

2005. In 2020, ER has been ranked as 9th best Russian journal among all disciplines and 3rd in economics according to the Science Index used by the Russian Science Citation Index[1]. It has been over 10 years since Economy of Regions entered the global stage (it joined Scopus in 2010) and this is a good time to look at its main topics and how they have changed over time. Therefore, our research questions are:

(1) What are the main research directions (topics) presented in ER?

(2) How did the popularity of these topics change over time?

(3) Which topics have the highest number of citations, prevalence among manuscripts in Russian and English language, number of authors with foreign affiliation and third-party funding?

The rest of this paper is organised as follows. In the Data and Methodology section, we describe our textual and quantitative data, explain how topic modelling works and discuss its benefits. In the Results section, we define the topics, evaluate their popularity over time, and associate them with the number of citations, presence of authors with foreign affiliation, third-party funding, and language of the article. The last section concludes.

Data and Methodology

This study is based on 1307 articles published in ER between 2010 and 2021. The data was retrieved from Scopus abstract & citation database[2] on 3 February 2022[3]. As a result, we obtained the distribution of text documents over time demonstrated in Figure 1. The number of publications peaked at 138 a year after being included in the Scopus database, and then averaged around 103 publications per year.

As expected, scientists with affiliation in Russia participated in the vast majority of articles (1008 cases). However, authors who published in the journal are also affiliated in many other countries. Thus, apart from Russia, the countries with the largest presence of researchers in the journal are Kazakhstan and Italy (see Fig. 2).

Text Data Description

To have high accuracy of our textual analysis, we collected all the textual information available. Thus, the text-based information for each paper consists of the title of the article, its keywords and abstract. Some authors try to use different wording in title and keywords, and this will help the topic model to better identify the most important words in each document. Given the original texts, we have 10458 unique tokens and 334916 word occurrences in total. The average number of words per document is 254.

A mandatory step before building a topic model is preliminary processing of textual data. We used the standard data cleansing steps described in recent literature (Aggarwal, 2018; Uglanova, Gius, 2020; Savin et al., 2020; Savin, Ott, Konop, 2021). In particular, the text documents were divided into separate elements (tokens), which were subsequently cleared of punctuation and stop words, and also brought to a

[1] Comparison of bibliometric indicators of Russian scientific journals. eLibrary.ru. Retrieved from: https://elibrary.ru/titles_compare.asp (Date of access: 20.02.2022).

[2] About Scopus — Abstract and citation database. Elsevier. Retrieved from: https://www.elsevier.com/solutions/scopus (Date of access: 20.02.2022).

[3] We do not consider publications in Economy of Regions until 2010, since the necessary data (e.g. the number of citations) is not available for the earlier period.

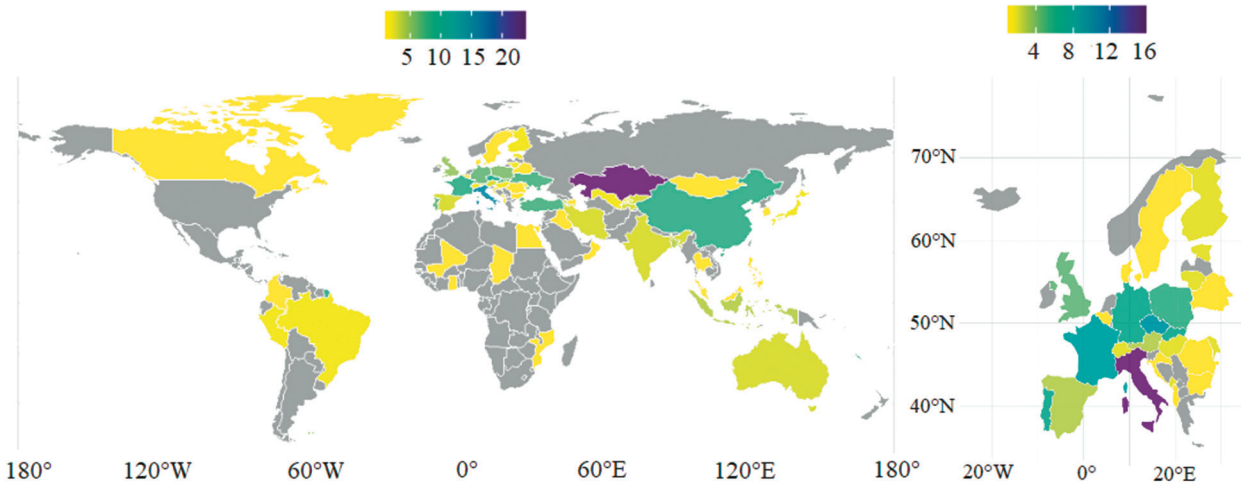

Fig. 2. *International participation in ER journal*
Note: The left heatmap shows international participation from the whole world, while the right one demonstrates affiliations in European countries

Table 1

Descriptive statistics of quantitative variables

Variable	Min	Max	Median	Mean	StDev
Publication year	2010,00	2021,00	2015,00	2015,24	3.47
Number of authors	1.00	5.00	2.00	2.17	0.99
Number of citations	0.00	49.00	1.00	2.45	3.66
English language	0.00	1.00	0.00	0.21	0.41
Author with foreign affiliation	0.00	1.00	0.00	0.28	0.45
Third-party funding	0.00	1.00	0.00	0.20	0.40

dictionary form (for example, replacing "challenging" with "challenge") using Wordnet-based lemmatization engine with part-of-speech tagging (Fellbaum, 2005; Voutilainen, 2003). We also removed words that are too rare (i. e., that appear less or equal to 5 times in all the documents). Finally, we have identified stable word sequences called bigrams (e.g., "gross_product", "european_union"). As a result, our final dataset contains 2583 unique words for building a topic model and 154 850 total word occurrences. The average number of words per document after pre-processing fell to 118.

Quantitative Data Description

We also use quantitative variables in our study, which are provided in the Scopus database as document-level metadata. The set of quantitative variables used and their descriptive statistics are presented in Table 1. This data covers two needs at the same time. Firstly, it helps topic model to extract topics from documents more accurately (He et al., 2009; Speier, Ong, Arnold, 2016; see also our description of the Structural Topic Modelling method). And, secondly, the data will be used later to see which topics tend to have more citations, authors with foreign affiliation, third-party funding and are written in English.

Note here that we do not distinguish between one or multiple authors with foreign affiliation, neither whether the author with foreign affiliation is the sole author. The variable (author with foreign affiliation) is binary and should be interpreted as presence of an author with foreign affiliation. Similarly, the variable on third-party funding is also binary and does not distinguish whether one or multiple grants have been received. We made these two simplifications since distinguishing these variables further would complicate both data preprocessing and its analysis.

It is also worth stressing that the covariates we use do not have high Pearson correlation coefficients, but those coefficients that we present below are significant at the 1 % level. The highest correlation is between year of publication and presence of third-party funding (0.5), which reveals the fact that studies published in ER in recent years have been supported by either Russian (Russian Science Foundation, Russian Foundation for Basic Research) or sometimes also foreign grants (e.g., Ministry of Education and Science of the Republic of Kazakhstan or National Science Centre Poland). Affiliation of one of the co-authors to a foreign university, contrary to our initial expectation, is not so strongly correlated with

English language of the article (coefficient is 0.44, which is a moderate value). This can be explained by a considerable share of native Russian-speaking authors who went abroad but continue collaboration with their colleagues in Russia and publish in Russian. Finally, the share of articles with an author affiliated abroad has a slight tendency to increase in ER over time (Pearson correlation with year of publication is 0.12).

Structural Topic Modelling

To reveal topics present in our textual data (ER articles), we use topic modelling (TM). In simple words, TM clusters words into topics based on the measure of their co-occurrence, i. e. how often any pair of words appears in the same text (Murakami et al., 2017). For example, if we see the word "labour" in a topic labelled "Labour Economics", it means that this word appears relatively more often in combination with other words from this topic. Formally speaking, using methods from Bayesian inference TM assumes that each word in our documents is generated through a two-step process. First, each article has its own distribution of topics, and a topic is randomly drawn from it. Second, each topic has its own word distribution, and a word is randomly selected from this distribution for the topic chosen earlier. Therefore, each document is a result of repeating these two steps sufficiently many times, while the articles typically have multiple topics in different proportions. TM discovers the topics by fitting this two-step model to replicate best the underlying data. Compared to simple count of keywords, TM has the advantage of considering words not in isolation, but accounting for their context, which can influence the meaning of the words.

An advantage of structural topic modelling (STM) over classical TM is that it includes additional information about the articles, like — in our case — the year of publication, English language, third-party funding, the presence of an author with foreign affiliation and the number of citations. Using additional data as covariates at the stage of estimating a topic model has proven to produce topics with higher predicting power and interpretability (Roberts et al., 2014). Furthermore, later we can statistically relate those covariates to the produced topics showing how the prevalence of topics was changing over time, or which topics had a larger presence of authors with foreign affiliation. We apply STM using the associated R package developed by Roberts, Stewart and Tingley (2019).

Results

Extracting and Describing Topics

Determining the optimal number of topics is a step to start building a topic model after preprocessing the text data. This is a difficult task, as it requires certain trade-offs in terms of maximising model performance. We follow some best practices for working with scientific texts and run the model for the number of topics from 3 to 30 to capture the diversity in their content (Blei, 2012; Savin, Drews, van den Bergh, 2021; Savin, Chukavina, Pushkarev, 2022). For each of the models, we document three metrics: heldout log-likelihood, exclusivity, and semantic coherence. They indicate (1) the predictive power of the model, (2) the degree of overlap between popular words within each topic, and (3) the degree of co-occurrence of words from the same topic in text documents, respectively.

In general, increasing the number of topics tends to increase the model's predictive power and topic exclusivity, but reduces their semantic coherence. Figure 3 demonstrates that 22 topics

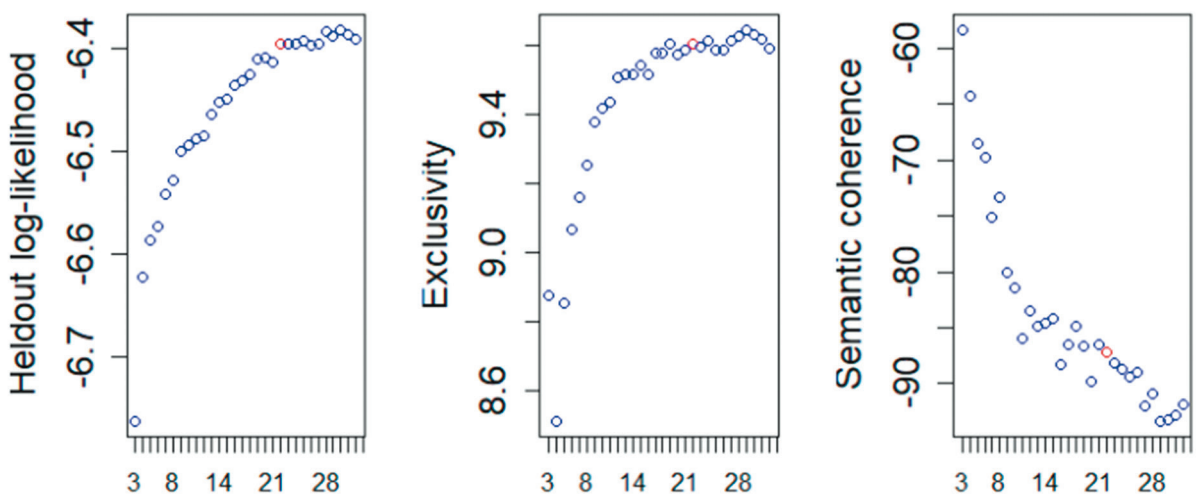

Fig. 3. *Model performance depending on the number of topics*

Fig. 4. *Word clouds for our 22 ER topics*

allow us to achieve close to the maximum possible predictive performance and exclusivity while maintaining semantic coherence at a not too low level. It is important to keep a relatively high semantic coherence of topics, as they must be unambiguously interpreted by human experts (Mimno et al., 2011).

Next, we turn to the description of the content on the topics we have chosen. We will start by presenting our topics as word clouds (see Figure 4), where word frequency is indicated by font size and exclusivity — by colour saturation. It is necessary when trying to understand the most important words in the topics. The word clouds show that the extracted topics are diverse. They represent combinations of words with more or less even distribution of weights among them instead of being highly concentrated around a few terms. For instance, topic 18 accentuates the words "human", "quality", "life", "health", "health_care" and "medical" to emphasise the problems in the efficiency of healthcare system and how this could potentially worsen the quality of life in Russian regions. A good example illustrating this topic is the article by Baskakova et al. (2020) studying the role of HIV infection on the quality of life in the Sverdlovsk region.

The labels of our topics (first introduced in Figure 4) have been chosen to reflect their main content as precisely as possible. To do this properly, one needs to coordinate the most frequent and exclusive words for each topic and original texts with the highest proportion of the respective topic — in our case, these are the titles, abstracts and keywords of the articles. For example, for the topic 10, the most important words "financial", "tax", "public" in combination with the illustrative title of the article "Sustainability of regional budget revenues and its sources"[1] suggest the topic label "Fiscal Policy". In Table 2 we provide 3 illustrative article titles for each topic[2].

It is worth stressing again that each topic is a set of words that co-occur in the same ER articles with high probability. Each ER article is a combination of topics with different prevalence (weights), and even words themselves can belong to different topics with different probabilities (see colour of words in Figure 4). Topics are formed solely based on how often words in our ER articles appear together. Therefore, it is likely that certain topic may have similar meaning but not form a single topic because authors of the articles where these topics dominate have preferred to use different wording (Liu, Nzige, Li, 2019). Good examples are T5 on economic geography and T13 on spatial economics. To better distinguish between the content of the topics, we took advantage of the illustrative papers mentioned in Table 2.

Exploring Topic Popularity over Time

Once the topics have been introduced, we proceed with analysing their popularity (i. e., relative shares[3]) across all documents. Moskaleva et al. (2018) in their article conducted a content analysis of scientific papers by constructing the distribution of RSCI articles by subject areas. We go further and trace the distribution of topics within one subject area, not only in general, but also over time. Figure 5 demonstrates a sorted list of 22 topics ranging from the most prevalent topic of industrial policy (T2) to the least prominent topic of digital economy (T17). Each of the three most frequent topics (T2, T9, T1) is almost twice as prevalent as the three rarest ones (T6, T7, T17). It is logical that topic of industrial policy is in the lead, because historically Sverdlovsk oblast and the Ural Federal District as a whole are pronounced industrial territories, while many studies in ER are focused on these territories (Kochetkov, Vuković, Kondyurina, 2021).

As mentioned above, one of the benefits of the STM model is the ability to use document-level covariates to improve the predictive power of a topic model. The same factors can be useful in explaining the variation of topic shares in our text documents (see "Quantitative data description" section). More precisely, we are trying to answer the question of which topics significantly increased their relative share over time and which ones have lost; and which topics have disproportionally more (less) citations, authors with foreign affiliation, third-party funding and are written in English or Russian language. For these purposes, we estimate the following linear model for each of the 22 topics (indexed with *i*):

$$Topic\ Prevalence_i \sim Constant_i + Year + English\ language + Third\ party\ funding + Author\ with\ foreign\ affiliation + Number\ of\ authors + Number\ of\ citations + Residual_i \quad (1)$$

[1] Malkina, M. (2021). Sustainability of Regional Budget Revenues and Its Sources. Ekonomika regiona [Economy of Region], 17(4), 1376–1389. DOI: 10.17059/ekon.reg.2021-4-23.

[2] In reality, we went through 10 most popular and exclusive words for each topic and ten illustrative texts (titles, abstracts, keywords) to come up with optimal topic titles. But in this paper for brevity reasons, we present only three illustrative titles. Note that illustrative titles are titles chosen from the ten texts with the highest prevalence (weight) of these topics.

[3] The sum of all topic shares (for the respective period of time) should be equal to 1.

Table 2

Illustrative article titles for each of 22 extracted topics

Topic 1. Mathematical Methods & Models
«Dynamic optimization of the complex adaptive controlling by the structure of enterprise's product range»
«Multi-criterion optimization of production range generation by an enterprise»
«Dynamic model of minimax control over economic security state of the region in the presence of risks»
Topic 2. Industrial Policy
«Strategic vector of economic dynamics of an industrial region»
«Industrial policy: genesis, regional features and legislative provision»
«Conditions and factors of structural modernization of a regional industrial system»
Topic 3. Infrastructure Development
«The role of transport corridors in Timan-North Ural region mineral-raw-material base development»
«Problems of economic security in Russian transportation and intermediate carrier infrastructure»
«Region tourist and recreation complex development»
Topic 4. Agriculture
«Seven fundamental economic characteristics exclusivity of agrifood supply chains (part 1)»
«Involvement of rural households in solving the problems of import substitution»
«Using land resources in agriculture of Belgorod Oblast»
Topic 5. Economic Geography
«On the essence of the brand of territory»
«Vital stability of territory: the contents and ways of strengthening»
«The features of geo-ecological assessment within the geo-eco-socio-economic approach to the development of northern territories»
Topic 6. Labour Economics
«Criteria, probability and degree of instability of employment taking into account the features of the Russian labour market»
«Regulation of precarious employment in single-industry towns»
«Persons of pre-retirement age in the labour market: employment problems and support measures»
Topic 7. Firm Performance
«Internal sources to increase financing for fixed investments in a company»
«Russian gas companies' financial strategy considering sustainable growth»
«The vulnerability and resilience of the company in modern economic space»
Topic 8. Demographic & Migration Policy
«Using cohort fertility indicators to assess and predict the effectiveness of demographic policies»
«Factors of emigration from Russia: regional features»
«Current migration processes in the Far East (on the example of Jewish Autonomous Oblast)»
Topic 9. Regional Security
«The economic security of the Volga Federal District regions»
«Theoretical and methodological approaches to the diagnosis of the region's state material reservation system status»
«Region as a self-developing socio-economic system: crossing the crisis»
Topic 10. Fiscal Policy
«Consolidated taxation and its consequences for regional budgets»
«Public sector financial balances based on the system of national accounts»
«Failures of big business tax administration and their impact on regional budgets»
Topic 11. Energy Economics
«Cost-effective management of electricity transmission in an industrial region»
«Regional aspects of price-dependent management of expenditures on electric power»
«Prospects for energy demand management in Russian regions»
Topic 12. Economic Theory
«Features of the creative method of Karl Marx»
«Paradoxes of economic theories and politics»
«Institutions of scientific efficiency: organizations of the Middle Urals»
Topic 13. Spatial Economics
«Spatial spillover effect of transportation infrastructure on regional growth»
«The spread of the covid-19 pandemic in Russian regions in 2020: models and reality»
«Spillover effects of the Russian economy: regional specificity»

End Table to next page

End Table 2

Topic 14. Entrepreneurship
«Global corporations and smaller actors in textile business: European perspective»
«Participation of Kazakhstan pharmaceutical companies in global value chains»
«Stimulation of managers in regional enterprises»
Topic 15. Inequality & Poverty
«Social attitudes and regional inequalities»
«Inequality of Spanish household expenditure for the 2006–2016 period — are we converging?»
«Modelling the impact of sanctions on income inequality of population in the target countries»
Topic 16. Monetary Policy & Trade
«Are exports and imports asymmetrically cointegrated? Evidence from the emerging and growth-leading economies»
«The relationship between inflation and inflation uncertainty in Turkey»
«The role of monetary policy in comparative advantage and trade balance of capital-intensive industry in Indonesia»
Topic 17. Digital Economy
«Technology of accelerated knowledge transfer for anticipatory training of specialists in digital economy»
«Implementation of digital technologies in financial management»
«Post-industrial technologies in the economy of the north-west of Russia»
Topic 18. Healthcare & Quality of Life
«The impact of HIV infection on the population's quality of life in regions»
«Assessment of population and territory rehabilitation efficiency regarding radiation exposure»
«Air pollution and public health in a megalopolis: a case study of Moscow»
Topic 19. Education & Human Capital
"Quality assessment of online learning in regional higher education systems"
«Drivers for development in regional higher education»
«Role of school education in development of human capital»
Topic 20. Economic Integration
«Influence of intraregional integration processes on socio-economic situation of Central Asia region»
«Trends and economic assessment of integration processes at the metal market»
«Development trends of the Russian regions»
Topic 21. Economics of Innovation
«Innovative development problems of private sector of engineering and defense industrial complexes»
«Innovation competitiveness of the Russian regions»
«Problems forming innovative-technological image of Russian regions»
Topic 22. Environmental Economics
"Harm to the resources of traditional nature management and its economic evaluation"
"The economic assessment of harm to the Arctic ecosystems at the development of oil and gas resources"
«Economic damage caused by consequences of the environmental impact of mining complex»

In Figure 6, which shows how topic popularity changes over time, 9 out of 22 topics were confirmed to have a significant time trend with at least 5 % level.

The prevalence of the topic of spatial economics (T13) has increased the most over time (+17.61 %), starting from 0.88 % in 2010 and peaking at 18.49 % in 2021. At the same time, despite the steepest uptrend (significant at the 1 % level), the relative share of this topic grew gradually between 2001 and 2018, and the explosive growth of its popularity occurred in 2019–2021. Topic 21 on the economics of innovation lost its popularity the most (–14.58 %) over the entire observation period. The topic lost half of its prevalence (16.99 %) in 2011 compared to 2010, and then gradually continued to fade, reaching its minimum (2.42 %) in 2021. This downtrend is also significant at the 1 % level.

These changes reflect the process of the journal concentrating on its primary research focus lying in the area of regional development and not innovation economics. This transition may have different reasons. On the one hand, the editors may direct the journal to the scope focusing on regional economics. According to its present own aims and scope, the journal "provides a platform for dialogue on socio-economic processes occurring at regional levels ranging from local areas to individual countries and [...] covers topics of regional development, regional economic and social policies, regional demographics, territorial management, urban and rural development, resource management and regional infrastructure"[1]. On the other

[1] About the Journal. Economy of Regions. Retrieved from: https://www.economyofregions.org/ojs/index.php/er (Date of access: 25.02.2022).

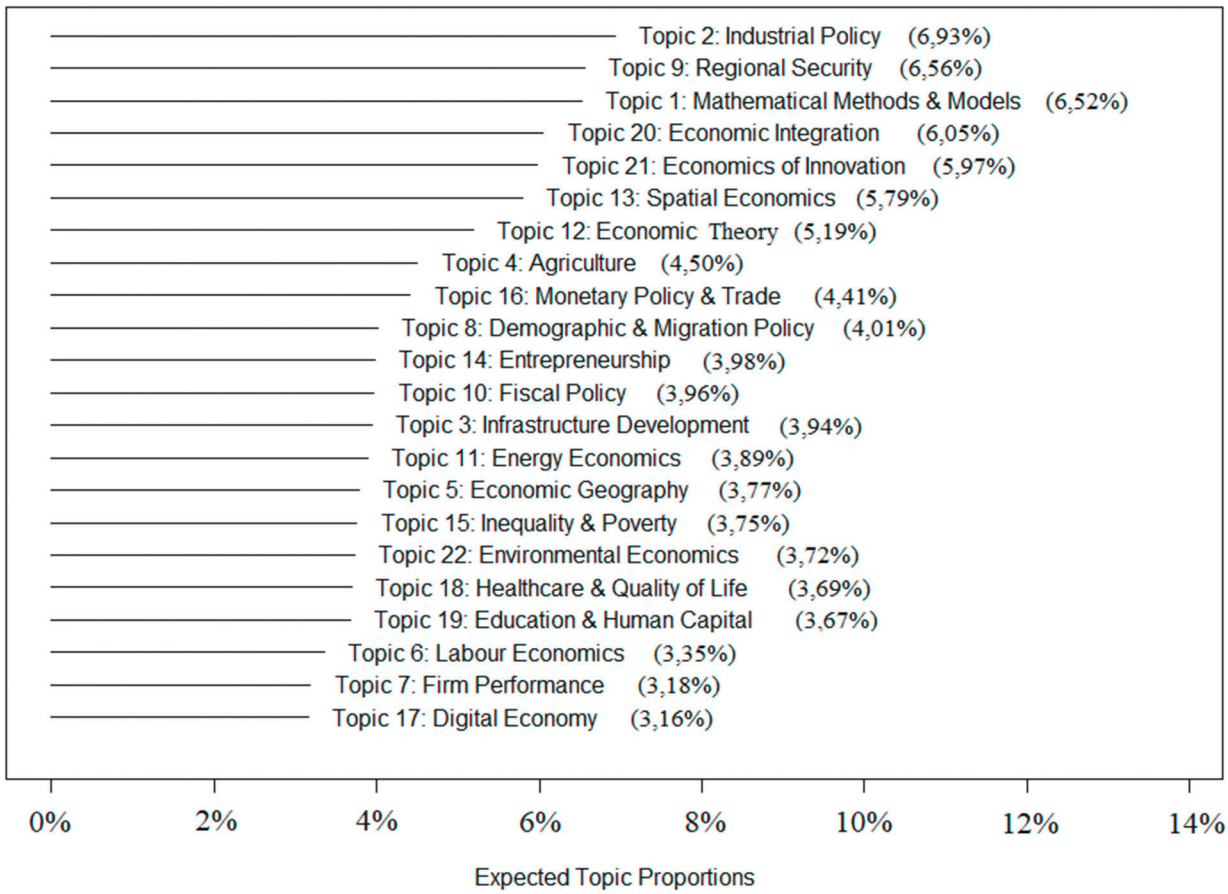

Fig. 5. *The prevalence of ER topics*

hand, researchers in the area of innovation economics may have changed their focus from ER to other journals. It is also worth mentioning that during the same time ER has considerably improved its quality indicators in Scopus from fourth to second quartile in subject areas like Social Sciences, Economics, and Geography, Planning and Development[1]. The latter fact demonstrates that the shift in research focus of the journal has been successful.

Among other topics, which have increased their prevalence over time (significant at the 1 % level), one can note T16 on monetary policy (+6.91 %) and T17 on digital economy (+2.86 %). The topic that has gradually lost in its share (significant at the 1 % level) is T9 on regional security (–3.48 %).

There are also areas of study that rank high in prevalence over time (Fig. 5) but did not exhibit significant up- or downward trends. These are T2 for industrial policy, which has an average prevalence of 6.9 %, and T1 on mathematical methods, with an average prevalence 6.5 %.

Establishing Statistical Relationship between Topic Prevalence and other Numerical Covariates

Finally, we analyse how research directions are related to the language of the article, authors with foreign affiliation, third-party funding, or the number of citations. We start by depicting the number of articles' distribution among topics with a grouping language attribute (see top panel in Figure 7). In particular, the bar chart shows what proportion of manuscripts in the corresponding language belongs to each topic. For example, about 7 % of articles in Russian and 5 % in English belong to topic 1[2].

The most popular topics of articles in Russian are T9 (Regional Security, 7.22 %), T2 (Industrial Policy, 7.13 %) and T1 (Mathematical Methods & Models, 6.82 %). At the same time, in English, the authors prefer to write about Monetary Policy & Trade (T16, 9.41 %) and Inequality & Poverty (T15, 8.32 %). For example, the most cited paper published in English in ER titled "Impact of

[1] Economy of Region. Scimago Journal & Country Rank. Retrieved from: https://www.scimagojr.com/journalsearch.php?q=21100242818&tip=sid&clean=0 (Date of access: 26.02.2022).

[2] Note that Figure 7 is constructed in a way to show how papers belonging to one of the two binary groups (Russian/English language, with or without third-party funding, with or without author with a foreign affiliation) are distributed among the 22 resulting topics. In other words, bars of each color sum up to 100 %.

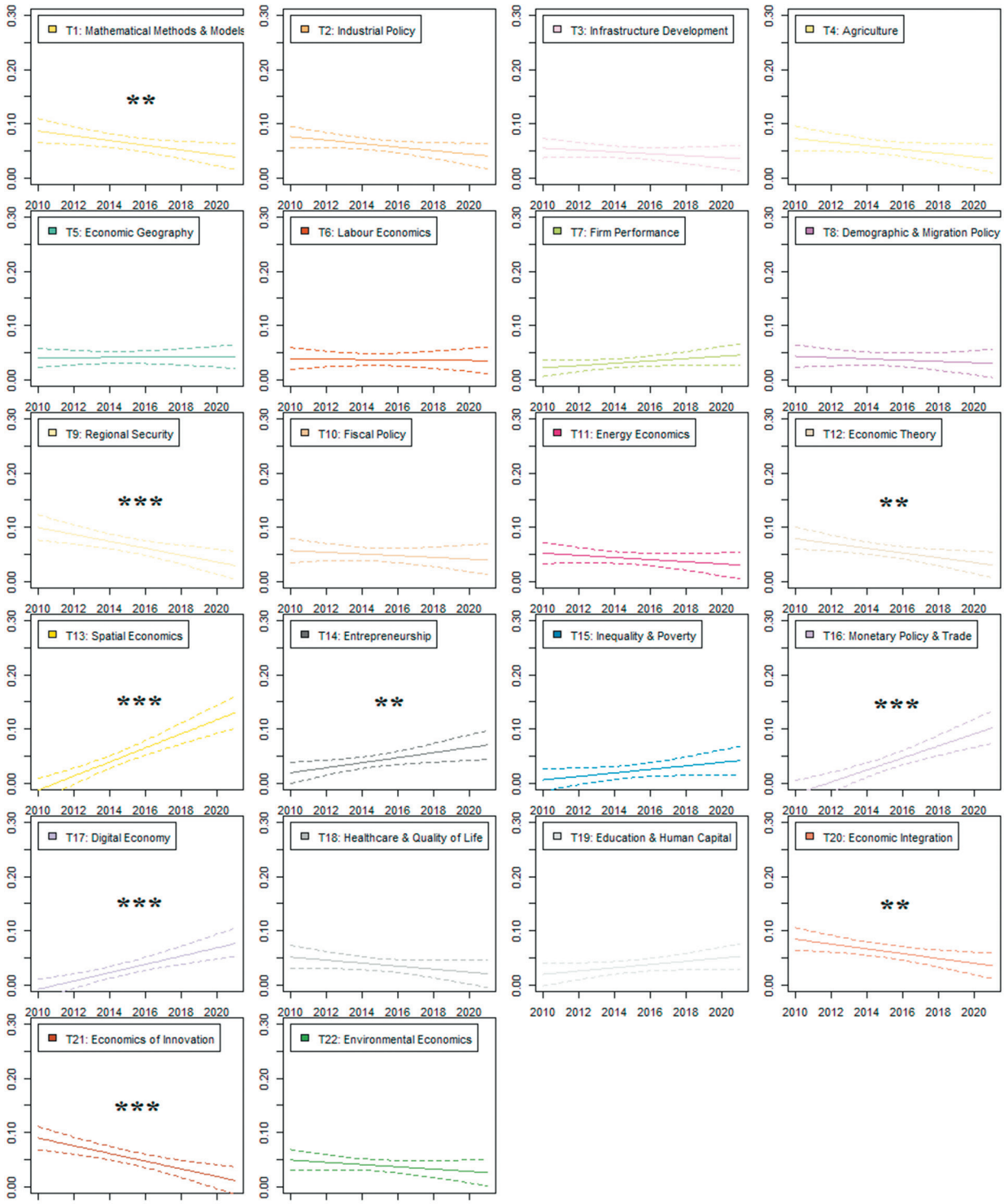

Fig. 6. *Changes in topic prevalence over time in ER*

*Note: The chart shows estimates of the effects of time on topic prevalence. Confidence intervals indicate a 95 % uncertainty range and include both regression and measurement uncertainties associated with the STM model. *** and ** denote 1 and 5 significance level, respectively.*

Political Instability on Economic Growth in CEE Countries" is mostly a mixture of T15 (27.8 %) and T16 (29.5 %).

Further, we examine how articles with and without foreign affiliation were distributed by topic (i. e., which topics were most and least likely to be covered by foreign-affiliated authors). The results in the middle panel in Figure 7 indicate that the manuscripts written with the participation of foreign-affiliated authors most often belong to topics 15 and 16, which is consistent with the findings obtained for the language of the ar-

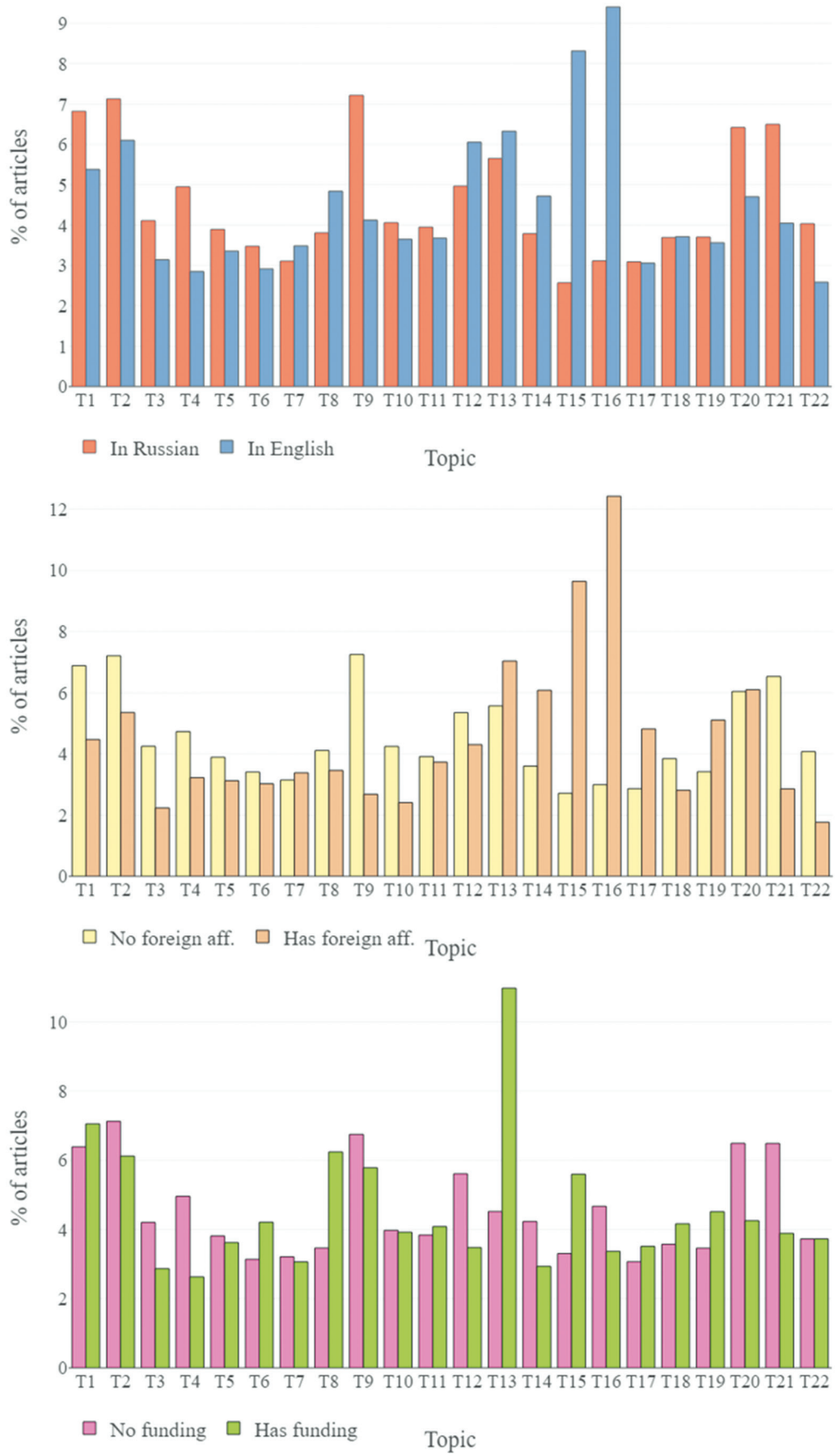

Fig. 7. *Distribution of manuscripts by topic in terms of language attribute, foreign affiliation and third-party funding*

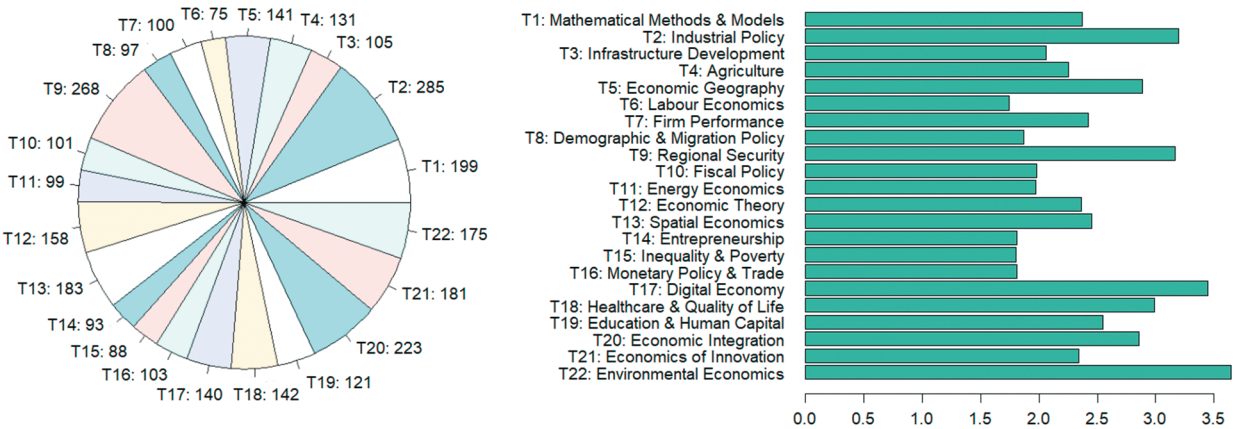

Fig. 8. *The total number of citations and the number of citations per article for each topic*

ticle. The article with the highest proportion of T16 (98.7 %) was written by an author affiliated in India, while the manuscript, which belongs to the T15 with 97.3 %, was created by researchers from Croatia (Sharma, Mittal, 2021; Muštra, Perovic, Golem, 2014).

We proceed with exploring how papers are distributed by topic depending on whether they are supported by third-party funding or not. In other words, the topics most and least likely to have received third-party funding will be identified. The bottom panel in Figure 7 shows the result. The most supported topic by a wide margin is T13 on spatial economics, accounting for about 11 % of funded manuscripts. For example, the work of Karim et al.[1] (belonging to 86.5 % to T13) on studying the spatial effect of transport infrastructure expansion on regional growth was funded by the Ministry of Research, Technology and Higher Education of Republic of Indonesia. The least funded topic is Agriculture (T4), although there is little difference from other less supported topics.

Now consider how the number of citations is distributed among topics. As expected, the topics with the highest prevalence of all time became the leaders in terms of the total number of citations. The three main topics in Figure 8 (left chart) are the same as the three main topics in Figure 5. T2 on industrial policy collected 285 citations, T9 on regional security received 268 citations and T1 on mathematical methods gathered 199 citations.

To be more objective in analysing the distribution of citations by topic, it is necessary to normalise citations by the number of articles published on the topic. As a result, we get the number of citations per article, which helps to eliminate the effect of varying topic popularity (right chart in Figure 8). Interestingly, topic 22 on environmental economics collected the most citations per article, which can be explained by the increasing concern about climate change and intensified discussion of climate policies (Drews, Savin, van den Bergh, 2022; van den Bergh et al., 2021). For example, the most cited article in ER so far by Leksin and Profiryev[2] (49 citations) discusses the impact of climate change on the Russian Arctic Zone and priorities for its sustainable development. Articles related to topics 2 on industrial policy, 9 on regional security and 17 on digital economy, on average, also collect more than 3 citations per article.

Conclusion

To conclude, we identified 22 topics published in ER over the last twelve years finding the topic of "Spatial Economics" (T13) was growing in the journal most rapidly, while the topic of "Economics of Innovation" (T21) was shrinking the most. At the same time, the largest topics presented in the journal are "Industrial Policy" (T2), "Regional Security" (T9) and "Mathematical Methods and Models" (T1). The topics written most frequently in English and/or by foreign-affiliated authors in ER are "Monetary Policy & Trade" (T16) and Inequality & Poverty (T15).

The most cited topics in the journal (per article) are "Environmental Economics" (T22) and "Digital Economy" (T17), which is not surprising given the urgency of climate change mitigation and the importance of the fourth industrial revolution. Finally, the topic that received proportionally more frequently third-party funding is T13 on "Spatial Economics".

Our results can serve as guidance for editorial board and contributors of the journal Economy of

[1] Karim, A., Suhartono & Prastyo, D. D. (2020). Spatial Spillover Effect of Transportation Infrastructure on Regional Growth. Ekonomika regiona [Economy of region], 16(3), 911–920. DOI: 10.17059/ekon.reg.2020–3-18.

[2] Leksin, V. & Profiryev, B. (2017). Socio-Economic Priorities for the Sustainable Development of Russian Arctic Macro-Region. Ekonomika regiona [Economy of region], 4(4), 985–1004. DOI: 10.17059/2017–4-2.

Regions, helping to identify topics that are going up or down in popularity (in terms of their share) or influence (in terms of citations and third-party funding) in ER. Our paper further illustrates the tools from computational linguistics like STM facilitating accurate and objective literature review analysis and providing complementary information and insights to traditional — and inevitably more subjective — human reviews (Savin, van den Bergh, 2021). The latter, however, is still necessary to address questions on contribution of ER articles to the literature on regional economics.

References

Aggarwal, C. C. (2018). Text Preparation and Similarity Computation. In: *Machine Learning for Text* (pp. 17–30). New York: Springer.

Ambrosino, A., Cedrini, M., Davis, J., Fioria, S., Guerzoni, M. & Nuccio, M. (2018). What topic modeling could reveal about the evolution of economics. *Journal of Economic Methodology, 25(4),* 329–348. DOI: 10.1080/1350178X.2018.1529215.

Asmussen, C. B. & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. *Journal of Big Data, 6(1),* 1–18. DOI: 10.1186/s40537–019–0255–7.

Baskakova, I. V., Podymova, A. S., Turgel, I. D. & Balandina, M. S. (2020). The Impact of HIV Infection on the Population's Quality of Life in Regions. *Ekonomika regiona [Economy of region], 16(1),* 114–126. DOI: 10.17059/2020–1-9. (In Russ.)

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM, 55(4),* 77–84. DOI: 10.1145/2133806.2133826.

Callaghan, M. W., Minx, J. C. & Forster, P. M. (2020). A topography of climate change research. *Nature Climate Change, 10(2),* 118–123. DOI: 10.1038/s41558 019 0684.

De Battisti, F., Ferrara, A. & Salini, S. (2015). A decade of research in statistics: a topic model approach. *Scientometrics, 103(2),* 413–433. DOI: 10.1007/s11192–015–1554–1.

Devitsyn, I. N. & Savin, I. V. (2020). Research Community Analytic Tool Based on Topic Modeling and Network Analysis. *Uspekhi kibernetiki [Russian Journal of Cybernetics], 1(4),* 13–21. DOI: 10.51790/2712–9942–2020–1-4–2. (In Russ.)

Drews, S., Savin, I. & van den Bergh, J. (2022). Climate change concern and policy support before and after COVID-19. *Ecological Economics, 199,* 107507. DOI:10.1016/j.ecolecon.2022.107507.

Fellbaum, C. D. (2005). WordNet and wordnets. In: K. Brown (Ed.), *Encyclopedia of Language and Linguistics, Second Edition* (pp. 665–670). Oxford: Elsevier.

Griffith, T. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101,* 5228–5235. DOI: 10.1073/PNAS.0307752101.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P. & Giles, C. L. (2009). Detecting Topic Evolution in Scientific Literature: How Can Citations Help? In: *CIKM '09: Proceedings of the 18th ACM conference on Information and knowledge management* (pp. 957–966). Hong Kong, China. DOI: 10.1145/1645953.1646076.

Kochetkov, D. M., Vuković, D. B. & Kondyurina, E. A. (2021). Challenges in Developing Urban Marketing Strategies: Evidence from Ekaterinburg. *Ekonomika regiona [Economy of regions], 17(4),* 1137–1150. DOI: 10.17059/ekon.reg.2021–4-7.

Liu, G., Nzige, J. H. & Li, K. (2019). Trending topics and themes in offsite construction (OSC) research: The application of topic modelling. *Construction Innovation, 19(3),* 343–366. DOI: 10.1108/CI-03–2018–0013.

Lüdering, J. & Winker, P. (2016). Forward or backward looking? The economic discourse and the observed reality. *Jahrbuecher fuer Nationaloekonomie und Statistik [Journal of Economics and Statistics] 236(4),* 483–515.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., … Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures, 12(2–3),* 93–118. DOI: 10.1080/19312458.2018.1430754.

Mimno, D., Wallach, H. M., Talley, E. M., Leenders, M. & McCallum, A. (2011). Optimizing Semantic Coherence in Topic Models. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing* (pp. 262–272). Edinburgh, Scotland.

Mo, Y., Kontonatsios, G. & Ananiadou, S. (2015). Supporting systematic reviews using LDA based document representations, *Systematic Reviews, 4(1),* 172. DOI: 10.1186/s13643–015–0117–0.

Moskaleva, O., Pislyakov, V., Sterligov, I., Akoev, M. A. & Shabanova, S. (2018). Russian Index of Science Citation: Overview and review. *Scientometrics, 116(1),* 449–462. DOI: 10.1007/s11192–018–2758-y.

Murakami, A., Thompson, P., Hunston, S. & Vajn, D. (2017). 'What is this corpus about?': using topic modelling to explore a specialised corpus. *Corpora, 12(2),* 243–277. DOI: 10.3366/cor.2017.0118.

Muštra, V., Perovic, L. M. & Golem, S. (2014). Social attitudes and regional inequalities. *Ekonomika regiona [Economy of region], 1,* 66–73. DOI: 10.17059/2014–1-5.

Roberts, M. E., Stewart, B. M. & Tingley, D. (2019). STM: An R Package for Structural Topic Models. *Journal of Statistical Software, 91(2),* 1–40. DOI: 10.18637/jss.v091.i02.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., … Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58(4),* 1064–1082. DOI: 10.1111/ajps.12103.

Savin, I. & van den Bergh, J. (2021). Main topics in EIST during its first decade: A computational linguistic analysis. *Environmental Innovation and Societal Transition, 41,* 10–17. DOI:10.1016/j.eist.2021.06.006.

Savin, I., Chukavina, K. & Pushkarev, A. (2022). Topic-based classification and identification of global trends for startup companies. *Small Business Economics*. DOI: 10.1007/s11187–022–00609–6.

Savin, I., Drews, S. & van den Bergh, J. (2021). Free associations of citizens and scientists with (green) economic growth: A computational linguistics analysis. *Ecological Economics, 180,* 106878. DOI: 10.1016/j.ecolecon.2020.106878.

Savin, I., Drews, S., Mestre Andrés, S. & van den Bergh, J. (2020). Public views on carbon taxation and its fairness: A computational linguistics analysis. *Climatic Change, 162,* 2107–2138. DOI: 10.1007/s10584 020 02842.

Savin, I., Ott, I. & Konop, C. (2021). Tracing the evolution of service robotics: Insights from a topic modeling approach. *Technological Forecasting and Social Change, 174,* 121280. DOI: 10.1016/j.techfore.2021.121280.

Sharma, V. & Mittal, A. (2021). Revisiting the Dynamics of the Fiscal Deficit and Inflation in India: the Nonlinear Autoregressive Distributed Lag Approach. *Ekonomika regiona [Economy of Region], 17(1),* 318–328. DOI: 10.17059/EKON.REG.2021–1-24.

Speier, W., Ong, M. K. & Arnold, C. W. (2016). Using phrases and document metadata to improve topic modeling of clinical reports. *Journal of Biomedical Informatics, 61,* 260–266. DOI: 10.1016/j.jbi.2016.04.005.

Uglanova, I. & Gius, E. (2020). The Order of Things. A Study on Topic Modelling of Literary Texts. In: CHR 2020: *Workshop on Computational Humanities Research* (pp. 57–76). Amsterdam, The Netherlands.

van den Bergh, J., Castro, J., Drews, S., Exadaktylos, F., Foramitti, J., Klein F., … Savin, I. (2021). Designing an Effective Climate-Policy Mix: Accounting for Instrument Synergy. *Climate Policy, 21(6),* 745–764. DOI: 10.1080/14693062.2021.1907276.

Voutilainen, A. (2003). Part-of-Speech Tagging. In: *R. Mitkov (Ed.), The Oxford handbook of computational linguistics* (pp. 219–232). New York: Oxford University Press Inc.

About the Authors

Ivan V. Savin — Doctor of Science (Econ.), Professor, Academic Department of Economics, Ural Federal University; Researcher, Institute of Environmental Science and Technology (ICTA), Universitat Autonoma de Barcelona; Scopus Author ID: 55539536700; https://orcid.org/0000-0002-9469-0510 (19, Mira St., Ekaterinburg, 620002, Russian Federation; Cerdanyola del Vallès, Barcelona, Spain; e-mail: ivan.savin@uab.cat).

Nikita S. Teplyakov — PhD Student, Research Assistant, Ural Federal University; Scopus Author ID: 57219122917; https://orcid.org/0000-0003-2522-8207 (19, Mira St., Ekaterinburg, 620002, Russian Federation; e-mail: nekit_teplykov@mail.ru).

Информация об авторах

Савин Иван Валерьевич — доктор экономических наук, профессор кафедры экономики, Уральский Федеральный университет им. первого Президента России Б. Н. Ельцина; исследователь, Институт наук и технологий об окружающей среде, Автономный университет Барселоны; Scopus Author ID: 55539536700; https://orcid.org/0000-0002-9469-0510 (Российская Федерация, 620002, г. Екатеринбург, ул. Мира, 19; Испания, г. Барселона, Серданьола дель Вайес; e-mail: ivan.savin@uab.cat).

Тепляков Никита Сергеевич — аспирант, лаборант-исследователь, Уральский Федеральный университет им. первого Президента России Б. Н. Ельцина; Scopus Author ID: 57219122917; https://orcid.org/0000-0003-2522-8207 (Российская Федерация, 620002, г. Екатеринбург, ул. Мира, 19; e-mail: nekit_teplykov@mail.ru).

Дата поступления рукописи: 10.03.2022.
Прошла рецензирование: 06.04.2022.
Принято решение о публикации: 07.04.2022.
Received: 10 Mar 2022.
Reviewed: 06 Apr 2022.
Accepted: 07 Apr 2022.