




Domain-specific Expertise in Economics, Business and Finance of Research Institutions in Poland

ANNA STASZEWSKA-BYSTROVA

corresponding author

Chair of Econometric Models and Forecasts, Faculty of Economics and Sociology, University of Lodz, ul. Rewolucji 1905 r. 37/39, 90-214 Łódź

✉ anna.bystrova@uni.lodz.pl

 ORCID: 0000-0002-3941-4986

VICTOR BYSTROV

Department of Economic Mechanisms, Faculty of Economics and Sociology, University of Lodz

✉ victor.bystrov@uni.lodz.pl

 ORCID: 0000-0003-0980-2790

Abstract

Motivation: The efficacy of research institutions, including those active in the areas of economics, business and finance, is often measured by published high-quality scientific articles. Whether the scientific output depends on thematic specialization is an open question. Scientific profiles can be estimated based on the publications affiliated with the research entities using the topic modelling techniques. The results can be used to compare the research mixes of different institutions and to measure the degree of their specialization.

Aim: We apply the latent Dirichlet allocation model (LDA) to evaluate research diversity in the fields of economics, business and finance of major scientific institutions in Poland. The importance of various research areas is evaluated using two metrics.



Results: The obtained rankings of topics provide information on the distribution of expert knowledge and specialization of various research entities in Poland.

Keywords: topic models; text analysis; latent Dirichlet allocation; research topics; research expertise

JEL: C49; O3; O43

1. Introduction

The main aim of this paper is to evaluate research diversity in the areas of economics, business and finance of major research institutions in Poland. The analysis is based on the latent Dirichlet Allocation (LDA) model of Blei et al. (2003), which allows revealing topics discussed in journal articles authored by researchers with Polish affiliations and analysing the prevalence of these topics among research institutions. The results provide insights into the domain-specific competence and interests of researchers from different organizations.

We focus on papers published in journals indexed in the Web of Science Core Collection in the years 2000-2024. The analysis is performed using abstracts and bibliographic information for these articles. Since research topics are not pre-defined but are treated as latent in the LDA approach, in the first step, we estimate the topics discussed in the collected corpus. The estimation results include information on the weights of each topic in every paper, providing a deeper insight into contents and focus of the publications than, e.g., typical keywords or classification codes. In the second step, we use these weights to assess the importance of the uncovered research topics for each institution.

It can be expected that institutions under study differ with respect to their research mixes, as regards the types of dominant topics as well as their weights. Such research diversity is likely to stem from various missions and objectives of some of these institutions, which include universities, business and economic schools of higher education, the academy of sciences, a university of technology and the central bank. Some other factors which might also have an impact on narrower or wider research areas are the size of an institution, its location, and past activities and interests of researchers employed there.

Our analysis is related to the literature on topic modelling in the following ways. Firstly, we provide a novel application of the LDA model to textual data from the fields of economics, business and finance. Some previous analyses of abstracts, articles and other sources include, for example, the papers by Mishra et al. (2024) who studied evolution of topics within computational economics or Bystrov et al. (2024a) who considered topics of economics articles from Poland and Germany to investigate the links between topics' importance and economic developments. Using abstracts from finance articles, Aziz et al. (2022) investigated the main topics of research using machine



learning methods and Liu and Wang (2024) studied thematic evolution of this strand of literature. Applications to business data include e.g. the paper by Bastani et al. (2019) who used LDA to analyse consumer complaints.

Secondly, we evaluate whether research topics can be properly inferred from scientific abstracts. This question was also studied e.g. by Syed and Spruit (2017) who compared coherence of LDA topics obtained on the basis of full-text data to those derived from abstracts. The authors concluded that analyses based on abstracts may capture broader topics and can include more noise terms. However, these effects become insignificant for larger samples of documents (see also Cao et al., 2023). Since our corpus includes more than 7000 documents and we obtain topics that are fully interpretable, we conclude that abstract data are, in our case, informative to a satisfactory degree about the latent research topics.

The structure of the paper is as follows. Section 2 presents research institutions included in our study. In section 3 we describe the source, structure and preparation of textual data used in the analysis. Section 4 provides a description of the methods for modelling latent research topics and computing their prevalence among different affiliations. The main results are presented in section 5. Section 6 contains the conclusions and perspectives for future research.

2. Research institutions and publishing policy

In this subsection we provide an overview of research institutions in Poland and the factors that influence their publication policy and output. In the description we focus on 16 research organizations analyzed in the paper. The selection criteria which led to obtaining the particular set of entities are described in detail in section 3.

Most of research organizations from our sample (14 out of 16) are institutions of higher education and science that can be characterized using disciplines in which scientific activities are conducted. There are seven universities (University of Warsaw, University of Lodz, Nicolaus Copernicus University, University of Szczecin, Jagiellonian University, University of Warmia and Mazury and University of Gdansk), six higher schools of business and economics (Warsaw School of Economics, Cracow University of Economics, Poznan University of Economics and Business, Wroclaw University of Economics and Business, Kozminski University and University of Economics in Katowice) and one university of technology (Gdansk University of Technology). All these entities apart from Kozminski University are public. The sample also contains Narodowy Bank Polski and Polish Academy of Sciences. To briefly describe research conducted by these institutions, for which disciplines are not available, we use the information from their webpages.

Table 1 provides information from the POL-on integrated system of information on science and higher education on scientific disciplines from

the area of social sciences of the institutions from our sample. We consider disciplines that are closely related to research fields from the Web of Science analyzed in this paper. The table also presents scientific categories assigned to the institutions by the Minister of Science and Higher Education within individual scientific disciplines for the evaluation period 2017-2021. Evaluation provides an assessment of the quality of scientific activity which includes: publications, grants and the impact on the functioning of society and the economy. The highest scientific category that can be awarded in each discipline is A+ and the lowest is C. The category has bearing on the rights to conduct studies, doctoral schools, award degrees and titles and the amount of funds received from the state budget. For the evaluation purposes, publication activities are assessed based on the list of scientific journals and peer-reviewed proceedings of international conferences published by the Ministry of Science and Higher Education which takes into account international rankings of journals as well as evaluation by the national experts. In our study, we consider publications from Web of Science which provides an external source of information on the quality of academic journals.

As follows from Table 1 there is a distinction between universities and most of the remaining science and higher education institutions (schools of business and economics and the university of technology), which consists in scientific activities in a broader scope of disciplines including economics and finance, socio-economic geography and spatial management, political and administrative sciences and management and quality studies. Only one higher school of economics, namely Cracow University of Economics, is also active and evaluated in all these fields. Warsaw School of Economics is not involved in research of the socio-economic geography and spatial management, while Poznan University of Economics and Business, Wroclaw University of Economics and Business, Kozminski University, University of Economics in Katowice and Gdansk University of Technology are active in two disciplines: economics and finance and management and quality studies. All categories for every discipline and institution are A or B+ confirming high scientific standing in Poland of organizations selected for this study.

Table 1. Scientific activities and evaluation of research institutions

Name of institution	Discipline (awarded category)
University of Warsaw	economics and finance (A), socio-economic geography and spatial management (A), political and administrative sciences (B+), management and quality studies (A)
Warsaw School of Economics	economics and finance (B+), political and administrative sciences (B+), management and quality studies (B+)
University of Lodz	economics and finance (B+), socio-economic geography and spatial management (B+), political and administrative sciences (B+), management and quality studies (B+)



Name of institution	Discipline (awarded category)
Poznan University of Economics and Business	economics and finance (B+), management and quality studies (B+)
Cracow University of Economics	economics and finance (B+), socio-economic geography and spatial management (B+), political and administrative sciences (A), management and quality studies (B+)
Wroclaw University of Economics and Business	economics and finance (B+), management and quality studies (B+)
Kozminski University	economics and finance (B+), management and quality studies (A)
Nicolaus Copernicus University	economics and finance (A), socio-economic geography and spatial management (B+), political and administrative sciences (A), management and quality studies (A)
University of Szczecin	economics and finance (B+), socio-economic geography and spatial management (B+), political and administrative sciences (A), management and quality studies (A)
University of Economics in Katowice	economics and finance (B+), management and quality studies (B+)
Jagiellonian University	economics and finance (B+), socio-economic geography and spatial management (A), political and administrative sciences (A), management and quality studies (A)
Gdansk University of Technology	economics and finance (A), management and quality studies (A)
University of Warmia and Mazury	economics and finance (B+), socio-economic geography and spatial management (A), political and administrative sciences (B+), management and quality studies (B+)
University of Gdansk	economics and finance (B+), socio-economic geography and spatial management (B+), political and administrative sciences (B+), management and quality studies (B+)

Source: Own preparation

The research undertaken by employees of Narodowy Bank Polski supports the accomplishment of the main goal of this institution that is ensuring macroeconomic and financial stability. To this end, the primary focus of research activities is on application of various methodological approaches and tools to analyse monetary and financial stability.

The Polish Academy of Sciences conducts research in almost all fields. The main research topics of the Committees on Economic Sciences and on Financial Sciences of the Polish Academy of Sciences are quite broad and include: economic theory, socio-economic systems, market economy, public finance, the evolution of economic sciences, economic development, global economy and finance.

3. Data

3.1. Data source and structure

The data were collected from the Web of Science Core Collection platform. We selected publications by specifying the publication type as article, the

publication date as the window 2000-2024 and the country as Poland. The research area of interest was indicated by setting the Web of Science Categories to one of the following: Economics, Business & Economics, Economics & Finance or Business. Then, to consider institutions with relatively high publication output, we narrowed the search by selecting affiliations with at least 200 articles. Their list with the corresponding number of publications is provided in the Appendix in Table 2.

In the next step, we downloaded full records for the selected publications¹. This dataset included information on journal articles and papers published in books. In the latter case, no abstracts were often available and so we disregarded the records on book chapters from further analysis. The final dataset consisted of: the main textual data in the form of abstracts of articles and the corresponding metadata with bibliographic information including e.g. the names and affiliations of the authors, title of the paper, journal title, publication year etc. The size of this database was 7079. Figure 1 presents the number of papers associated with different affiliations and Figure 2 shows the total number of articles published in each year.

3.2. Preprocessing

The data were prepared for LDA modelling by applying several typical preprocessing operations (see e.g. Chai, 2023). The purpose of preprocessing is to reduce dimensionality of the analysis by excluding those words from the corpus that can be treated as relatively irrelevant and reducing different forms of a word to one single form, e.g. by means of lemmatization or stemming. Some common approaches to dimensionality reduction is to filter out language-specific stop words (commonly used words which appear in any collection of documents), corpus-specific stop words and infrequent words.

In this paper, we pre-processed the textual data by: lemmatization of words, removal of language-specific stop words, lowercasing, removal of punctuation and numbers. Additionally, we disregarded lemmas shorter than 3 characters. Preprocessing also involved removing the following corpus-specific stop words: aim, analyse, analysis, article, author, discuss, evidence, examine, find, impact, main, paper, present, purpose, report, research, result, show, study and suggest. These terms can be expected to appear in any abstract, but are not helpful in identifying latent research topics. We also removed rare terms defined as those used in fewer than 0.5% of all documents. As discussed by Bystrov et al. (2025) such vocabulary pruning is, on the one hand, not likely to lead to information loss and on the other hand, increases computational efficiency.

¹ Data were retrieved on February 4, 2025

The resulting dataset was then used to compute a matrix showing how often each of the terms appeared in each document (document-term (DTM) matrix). The matrix had dimensions 7079×1822.

4. Methods

In the first step, to identify research topics, we analysed the abstracts using the latent Dirichlet allocation model introduced by Blei et al. (2003). The model describes a mechanism of generating documents from a set of latent topics and provides a statistical framework for inferring these topics from a corpus of documents.

The first assumption in the LDA approach, is that each document in a text corpus is a distribution over some latent topics. According to the second assumption, each topic is a mixture of corpus vocabulary. Denoting the number of topics by K , the number of unique terms in a corpus by V and the number of documents by N , these assumptions can be presented using the following probability vectors:

- 1) $\theta_n = (\theta_{n,1}, \dots, \theta_{n,K})$, where $\theta_{n,k}$ is a weight of topic k in document n , $\sum_{k=1}^K \theta_{n,k} = 1$ and $\theta_{n,k} > 0$ for each document, indicating that each document contains all topics with positive probabilities. In the LDA method, the document-topic probabilities follow the Dirichlet distribution with a single concentration parameter α ,
- 2) $\beta_k = (\beta_{k,1}, \dots, \beta_{k,V})$, where $\beta_{k,v}$ is a weight of term v in topic k , $\sum_{v=1}^V \beta_{k,v} = 1$ and $\beta_{k,v} = 0$ for each topic $k \in \{1, \dots, K\}$, presenting each topic as a mixture of corpus vocabulary. Topic-term probabilities are also assumed to have the Dirichlet distribution with a parameter η .

Given the vectors θ_n and β_k , the generation of each document from a corpus can be described in two steps: 1) for each word slot in a document, a topic $k \in \{1, \dots, K\}$ is drawn according to probabilities θ_n and 2) a word from the selected topic is chosen based on the distribution β_k .

The model can be estimated using the variational EM algorithm of Blei et al. (2003) or Gibbs sampling described by Griffiths and Steyvers (2004). In this application we use the latter estimation method as implemented in the R environment in *topicmodels* package (Grün and Hornik, 2011). We select the Dirichlet prior similarly to Griffiths and Steyvers (2004), who considered $50/K$ and 0.1 in the case of α and η , respectively. We consider the values of $5/K$ and 0.1 since our dataset is much smaller and driven by considerably fewer topics.

The LDA estimation requires the user to specify the number of topics to be modelled (K). Since this number is typically unknown, several methods were proposed to estimate it (see e.g. Cao et al., 2009; Arun et al., 2010; Mimno et al., 2011; Deveaud et al., 2014; Lewis and Grossetti, 2022; Bystrov

et al., 2022). Some of these approaches were compared in a Monte Carlo study by Bystrov et al. (2024b) who demonstrated good performance of the singular Bayesian information criterion (sBIC) of Bystrov et al. (2022). Thus, in this paper we apply sBIC in the model selection process.

The results of the estimation are then used to assess the importance of topics for various research institutions. This analysis is based on vectors of the estimated document-topic probabilities, $\hat{\theta}_n$. The weights of the topics for all preselected research institutions are calculated as follows. First, we search the metadata for each document (abstract) to identify all affiliations that belong to the list from Appendix A and select vectors of estimated topic probabilities for documents affiliated with each institution. Then, for each affiliation and each topic, we compute the mean weights of the topic.

An alternative indicator of institution-specific topic weights proposed in this paper is based on the dominant topic, i.e. topic with the highest weight for each abstract. It is calculated as the share of documents from a given institution dominated by specific topics.

To provide a measure of dispersion of the prevalence of topics, which can be treated as a specialization indicator, we calculate standard deviations for the estimated topic probabilities for both metrics (mean topic weights and the dominant topic).

5. Results

The first step of the analysis is the selection of the model. The sBIC criterion indicated that the LDA with 11 topics would have the minimal generalization error in predicting probabilities of new documents (when searching in the interval from 5 to 30 topics). However, for the purpose of evaluating topic prevalence across institutions/affiliations in the corpus of published research articles, we decided to consider a slightly larger number of topics. Starting from 11, we augmented the number of topics by one as long as every topic belonged to the set made up of the top three most important topics for every institution. This criterion made it possible to focus on relevant topics only and not to stretch the number of topics too much as compared to the number of topics selected by sBIC. The final number of topics chosen in this way was 16.

The estimated topics are presented in Figure 3 in a typical form of word clouds. The font size in these graphs corresponds to the weight of the respective word for a specific topic. For example, “model”, “method” and “forecast” are the most important terms for topic three. Their respective probabilities were estimated at 0.0837, 0.0168 and 0.0150. The word clouds presented include 50 most important words for each topic. This means that the remaining 1772 terms whose weights were also estimated are not shown. The order in which the word clouds are presented in Figure 3 is random.

The uncovered topics are quite coherent. Thus, in the next step, they were labelled based on the top words, but also titles and abstracts of the documents with the highest weight of the respective topics. These concise titles are also provided in Figure 3. The extracted topics included: 1) entrepreneurship, 2) financial markets, 3) econometric modelling and forecasting, 4) innovations, 5) regional studies, 6) economic growth, 7) foreign trade, 8) labour market, 9) banking and finance, 10) income and inequality, 11) consumer behaviour, 12) monetary and fiscal policy, 13) European Union studies, 14) management, 15) public policy and sustainability and 16) institutional economics.

Table 3 presents three most important topics for each institution. The prevalence is measured using mean topic weights. The complete topic-importance distributions computed using this metric are depicted in Figure 4.

In general, according to mean weights, research mixes and top three research topics of the investigated institutions differ from each other. Nevertheless, there are also some similarities, e.g. for specific types of research organizations.

Table 3. Top topics for research institutions: mean weights

Research Institution	Top topics	Topic weights
University of Warsaw	3 – Econometric modelling & forecasting	0.1019
	10 – Income & inequality	0.0914
	8 – Labour market	0.0806
Warsaw School of Economics	3 – Econometric modelling & forecasting	0.1092
	12 – Monetary & fiscal policy	0.0971
	9 – Banking & finance	0.0750
University of Lodz	13 – European Union studies	0.0845
	12 – Monetary & fiscal policy	0.0765
	6 – Economic growth	0.0755
Poznan University of Economics and Business	2 – Financial markets	0.1700
	14 – Management	0.0754
	1 – Entrepreneurship	0.0714
Cracow University of Economics	3 – Econometric modelling & forecasting	0.0993
	14 – Management	0.0925
	1 – Entrepreneurship	0.0730
Wroclaw University of Economics and Business	14 – Management	0.1284
	16 – Institutional economics	0.0770
	1 – Entrepreneurship	0.0739



Research Institution	Top topics	Topic weights
Kozminski University	1 – Entrepreneurship	0.1510
	7 – Foreign trade	0.1072
	9 – Banking & finance	0.0885
Polish Academy of Sciences	15 – Public policy & sustainability	0.1008
	3 – Econometric modelling & forecasting	0.1003
	16 – Institutional economics	0.0943
Nicolaus Copernicus University	16 – Institutional economics	0.1015
	3 – Econometric modelling & forecasting	0.0888
	2 – Financial markets	0.0802
University of Szczecin	14 – Management	0.1005
	16 – Institutional economics	0.0855
	4 – Innovations	0.0847
Narodowy Bank Polski	12 – Monetary & fiscal policy	0.2087
	3 – Econometric modelling & forecasting	0.1213
	6 – Economic growth	0.0872
University of Economics in Katowice	1 – Entrepreneurship	0.1553
	4 – Innovations	0.0956
	14 – Management	0.0899
Jagiellonian University	14 – Management	0.1212
	1 – Entrepreneurship	0.1085
	16 – Institutional economics	0.0957
Gdansk University of Technology	1 – Entrepreneurship	0.0882
	6 – Economic growth	0.0865
	7 – Foreign trade	0.0860
University of Warmia and Mazury	5 – Regional studies	0.1604
	2 – Financial markets	0.0807
	14 – Management	0.0713
University of Gdansk	14 – Management	0.0954
	11 – Consumer behaviour	0.0871
	4 – Innovations	0.0759

Source: Own preparation

There seems to be more research diversity in the group of universities. The most important research areas for University of Warsaw, described by topics 3, 10 and 8, are econometric modelling and forecasting, income and

inequality and the labour market. Articles published by researchers affiliated with University of Lodz mainly focus on European Union studies (topic 13), monetary and fiscal policy (topic 12) and economic growth (topic 6), however the weights distribution is relatively flat for this institution. Researchers from Nicolaus Copernicus University are interested to the largest extent in institutional economics (topic 16), econometric modelling and forecasting (topic 3) and financial markets (topic 2), while the main research areas at the University of Szczecin are represented by topics on management (topic 14), institutional economics (topic 16) and innovations (topic 4). The profile of the Jagiellonian University resembles that of several business and economics schools described above, as the most important topics for this institution are: management, entrepreneurship and institutional economics. Researchers from University of Warmia and Mazury focus on regional studies, financial markets and management (topics 5, 2 and 14, correspondingly). The weight of regional studies is quite considerable, as it amounts to 0.1604. Finally, the top three research areas for the University of Gdansk are represented by management (topic 14), consumer behaviour (topic 11), and innovations (topic 4).

The remaining three institutions: Polish Academy of Sciences, Narodowy Bank Polski and Gdansk University of Technology are neither higher schools of business and economics nor universities. According to our results, Polish Academy of Sciences has the largest specialist knowledge in the areas of public policy and sustainability (topic 15), econometric modelling and forecasting (topic 3) and institutional economics (topic 16). As could be expected, the most important research areas for Narodowy Bank Polski (NBP) include monetary and fiscal policy (topic 12, most important) and economic growth (topic 6, third most important). The weight estimated for monetary and fiscal policy topic (0.2087) is the largest among all research areas and all institutions indicating a very strong focus on this topic by researchers from NBP. The second most prevalent research topic (econometric modelling and forecasting – topic 3) shows a strong quantitative component of papers affiliated with Narodowy Bank Polski. The main topics of articles affiliated with Gdansk University of Technology are entrepreneurship (topic 1), economic growth (topic 6) and foreign trade (topic 7).

Table 4 from the Appendix and Figure 5 present topic prevalence values for each institution measured using our second metric - the share of documents for which a specific topic had the highest weight (dominant topic). While Table 4 includes the top three indicator values, Figure 5 presents the complete distributions. By focusing on the top probabilities when constructing these indicators only, we do not take full advantage of the LDA estimation results. However, it might still be of interest to investigate the main areas of competence of researchers from various organizations. It can be seen that the two approaches do not order the topics' importance in exactly the same



way. For example, the top three research areas have the same ranking only for Cracow University of Economics. In the case of five further institutions (University of Warsaw, Warsaw School of Economics, Kozminski University, Polish Academy of Sciences and Nicolaus Copernicus University) the most important three topics are the same, however their ordering according to weights is different. For the remaining research organizations, apart from University of Lodz and Wroclaw University of Economics and Business, there is an overlap of two top research areas as measured by mean weights and the dominant weight. For University of Lodz and Wroclaw University of Economics and Business only one topic makes it to the top three in both rankings. According to the measure based on the dominant topic, articles affiliated with University of Lodz focus mainly on the monetary and fiscal policy (topic 12, ranked 2 by the mean weights metric), banking and finance (topic 9) and financial markets (topic 2). While the most important research area for Wroclaw University of Economics is still management (topic 14), institutional economics and entrepreneurship are replaced by innovations (topic 4) and public policy and sustainability (topic 15).

Table 5. Dispersion of topic prevalence

Research Institution	Dispersion (mean weights)	Dispersion (dominant topic)
University of Warsaw	0.0187	0.0313
Warsaw School of Economics	0.0190	0.0331
University of Lodz	0.0131	0.0209
Poznan University of Economics and Business	0.0308	0.0505
Cracow University of Economics	0.0171	0.0243
Wroclaw University of Economics and Business	0.0213	0.0283
Kozminski University	0.0318	0.0600
Polish Academy of Sciences	0.0198	0.0338
Nicolaus Copernicus University	0.0173	0.0276
University of Szczecin	0.0192	0.0277
Narodowy Bank Polski	0.0485	0.0798
University of Economics in Katowice	0.0326	0.0476
Jagiellonian University	0.0267	0.0401
Gdansk University of Technology	0.0185	0.0256
University of Warmia and Mazury	0.0295	0.0609
University of Gdansk	0.0166	0.0308

Source: Own preparation



In general, despite some differences, the shapes of the full distributions obtained using the two metrics remain similar, however as could be expected, the approach based on the dominant topic assigns even more weight to the topics which were identified as very important using mean probabilities.

Table 5 reports information on the dispersion of the topic prevalence for each research entity and both indicators of topic relevance. Higher values indicate focus on selected topics, while smaller values (meaning that probabilities for the topics were more similar) larger research diversity. No matter which metric is used, Narodowy Bank Polski is identified as most specialized and University of Lodz as most diversified.

Our results on scientific diversification are broadly in line with observations from section 2 which indicated differences among the institutions with respect to the number of disciplines in which scientific activities are conducted. In general, universities tend to be more diversified than higher schools of business and economics. However, some additional comments can be made. Firstly, Cracow University of Economics, which supports more disciplines than the remaining schools of business and economics, turned out to be more diversified also in view of our results. Secondly, researchers from Gdansk University of Technology (quite diversified according to our results) work on more subjects than could be expected based on the declared number of disciplines. Thirdly, Jagiellonian University and University of Warmia and Mazury are the least diversified universities in our sample.

As for the remaining institutions, the diversification of the Polish Academy of Sciences, which declares broad scope of research activities, is similar to that of universities from our sample. The largest specialization of Narodowy Bank Polski is in line with the declared, quite narrow, scope of research which aims to support the main aim of this institution.

6. Conclusions

The latent Dirichlet allocation model has become a standard tool for identifying unknown topics in publications from many areas. In this paper, we estimated the LDA model using abstracts of papers from the fields of economics, business and finance, affiliated with selected research institutions from Poland. The sixteen uncovered topics were coherent and interpretable, making it easy to label them.

The aim of the paper was to use the estimated document-topic probabilities provided by the LDA approach to evaluate research diversity of scientific institutions. We considered two alternative measures of topical research output that aggregate topic weights of research articles published by the analyzed institutions. The two measures can be used together as they provide slightly different ranking of research topics.

Our results inform about research mixes of the major Polish institutions carrying out research in the areas of economics, business and finance. In particular, we evaluated the importance of each of the research topics for every institution. By computing standard deviations of the estimated topic prevalence values, we also provided a measure of specialization/diversification for the research entities.

The main conclusions from our study are as follows. Firstly, there is a substantial diversity in the prevalence of research topics among major Polish institutions where studies in the areas of economics, business, and finance are conducted. Although there is a greater convergence of research interests for business schools than for universities and other types of institutions, even most similar entities have some distinctive features of their research profiles. Secondly, the analysed research institutions differ in terms of their degree of specialization. For some of the entities, there is no dominant topic (e.g. Polish Academy of Sciences and University of Lodz), while researchers from other institutions (e.g. Narodowy Bank Polski and Poznan University of Economics and Business) focus on a single topic to a much larger degree. When the whole distribution of topic probabilities is considered, the institutions can be ranked according to their specialization/diversification starting from Narodowy Bank Polski (most specialized) to the University of Lodz (most diversified).

Our results on diversification can be further explained by exploring the institutional settings for scientific activities in Poland. Most entities from our sample are institutions of higher education and science which are evaluated by the respective ministry within specific scientific disciplines. The awarded category depends to a large degree on the publications from these disciplines. The higher the category the more funds is channelled to the institution. Our results on larger diversification of universities than higher schools of business and economics partly reflect the general rule that universities conduct research in a larger number of scientific disciplines.

This work could be extended in several ways. One direction of future research is to use topic models with word embedding (see e.g. Grootendorst, 2022, Mu et al., 2024) which, in general, might provide topics which are more distinct from each other. At this point, it is not clear whether such an analysis could add to our results. Another interesting extension is to repeat our calculations using the full texts of the articles instead of abstracts. This could make it possible to extract more fine-grained topics and look into areas of specialization in more detail. However, collecting such a dataset would be more difficult and time-consuming. Further analyses could also take into account a different or extended set of publications than the Web of Science records used in this work. Finally, our results could be, perhaps, used to analyse whether research diversity or specialization are better drivers of scientific output.

References

- Arun, R., Suresh, V., Veni Madhavan, C.E. and Narasimha Murthy, M.N. (2010). On finding the natural number of topics with latent Dirichlet allocation: Some observations, in M.J. Zaki, J.X. Yu, B. Ravindran and V. Pudi (eds), *Advances in Knowledge Discovery and Data Mining*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 391–402.
- Aziz, S., Dowling, M., Hammami, H. and Piepenbrink, A. (2022). Machine learning in finance: A topic modeling approach, *European Financial Management* 28(3): 744–770.
- Bastani, K., Namavari, H. and Shaffer, J. (2019). Latent Dirichlet allocation (LDA) for topic modeling of the CFPB consumer complaints, *Expert Systems with Applications* 127: 256–271.
- Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet allocation, *Journal of Machine Learning Research* 3: 993–1022.
- Bystrov, V., Naboka-Krell, V., Staszewska-Bystrova, A. and Winker, P. (2025). Analysing the impact of removing infrequent terms on topic quality in LDA models, *Central European Journal of Economic Modelling and Econometrics*, 17: 61–85.
- Bystrov, V., Naboka-Krell, V., Staszewska-Bystrova, A. and Winker, P. (2024a). Comparing links between topic trends and economic indicators in the German and Polish academic literature, *Comparative Economic Research. Central and Eastern Europe* 2: 7–28.
- Bystrov, V., Naboka-Krell, V., Staszewska-Bystrova, A. and Winker, P. (2024b). Choosing the number of topics in LDA models – a Monte Carlo comparison of selection criteria, *Journal of Machine Learning Research* 25(79): 1–30.
- Bystrov, V., Naboka, V., Staszewska-Bystrova, A. and Winker, P. (2022). Cross-corpora comparisons of topics and topic trends, *Journal of Economics and Statistics* 242(4): 433–469.
- Cao, J., Xia, T., Li, J., Zhang, Y. and Tang, S. (2009). A density-based method for adaptive LDA model selection, *Neurocomputing* 72(7): 1775 – 1781.
- Cao, Q., Cheng, X. and Liao, S. (2023). A comparison study of topic modeling based literature analysis by using full texts and abstracts of scientific articles: a case of COVID-19 research, *Library Hi Tech* 41(2): 543–569.
- Chai, C. P. (2023). Comparison of text preprocessing methods, *Natural Language Engineering* 29(3): 509–553.
- Deveaud, R., SanJuan, E. and Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval, *Document numérique* 17(1): 61–84.
- Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics, *Proceedings of the National Academy of Sciences* 101 (suppl 1): 5228–5235.
- Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure, *arXiv preprint arXiv:2203.05794*.



- Grün, B. and Hornik, K. (2011). topicmodels: An R package for fitting topic models, *Journal of Statistical Software* 40: 1–30.
- Lewis, C. and Grossetti, F. (2022). A statistical approach for optimal topic model identification, *Journal of Machine Learning Research* 23: 1–20.
- Liu, P.-Y. and Wang, Z. (2024). Finance research over 40 years: What can we learn from machine learning?, *International Studies of Economics* 19(4): 472–507.
- Mimno, D., Wallach, H., Talley, E., Leenders, M. and McCallum, A. (2011). Optimizing semantic coherence in topic models, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Edinburgh, Scotland, UK., pp. 262–272.
- Mishra, M., Vishwakarma, S. K., Malviya, L. and Anjana, S. (2024). Temporal analysis of computational economics: a topic modeling approach, *International Journal of Data Science and Analytics* pp. 1–15.
- Mu, Y., Dong, C., Bontcheva, K. and Song, X. (2024). Large language models offer an alternative to the traditional approach of topic modelling, *arXiv preprint arXiv:2403.16248*.
- Syed, S. and Spruit, M. (2017). Full-text or abstract? Examining topic coherence scores using latent Dirichlet allocation, *2017 IEEE International conference on data science and advanced analytics (DSAA)*, IEEE, pp. 165–174.

Acknowledgements

Author contributions: authors have given an approval to the final version of the article. Author's total contribution to the manuscript: Anna Staszewska-Bystrova (50%); Victor Bystrov (50%).

Funding: financial support from the Polsko-Niemiecka Fundacja na rzecz Nauki (PNFN) (project no. 100-2024-00794) is gratefully acknowledged. The project also benefited from cooperation within HiTEC Cost Action CA 21163.

Appendix

Table 2. Articles by research institutions in the period 2000–2024

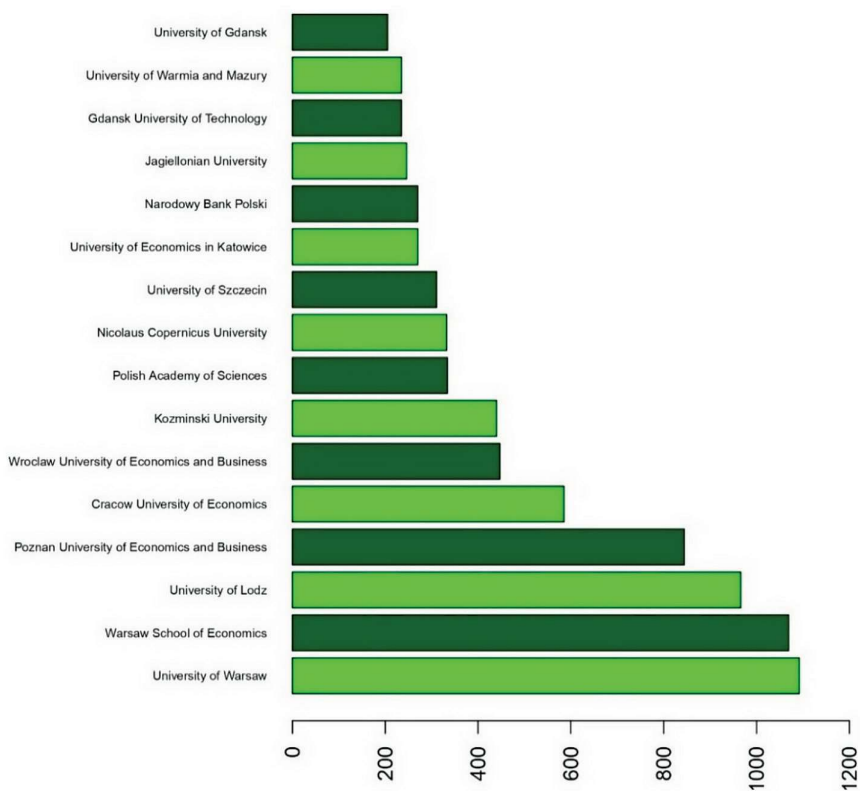
Research Institution	No. of articles
University of Warsaw	1092
Warsaw School of Economics	1069
University of Lodz	966
Poznan University of Economics and Business	844
Cracow University of Economics	585
Wroclaw University of Economics and Business	447



Research Institution	No. of articles
Kozminski University	440
Polish Academy of Sciences	334
Nicolaus Copernicus University	332
University of Szczecin	311
Narodowy Bank Polski	270
University of Economics in Katowice	270
Jagiellonian University	246
Gdansk University of Technology	235
University of Warmia and Mazury	235
University of Gdansk	205

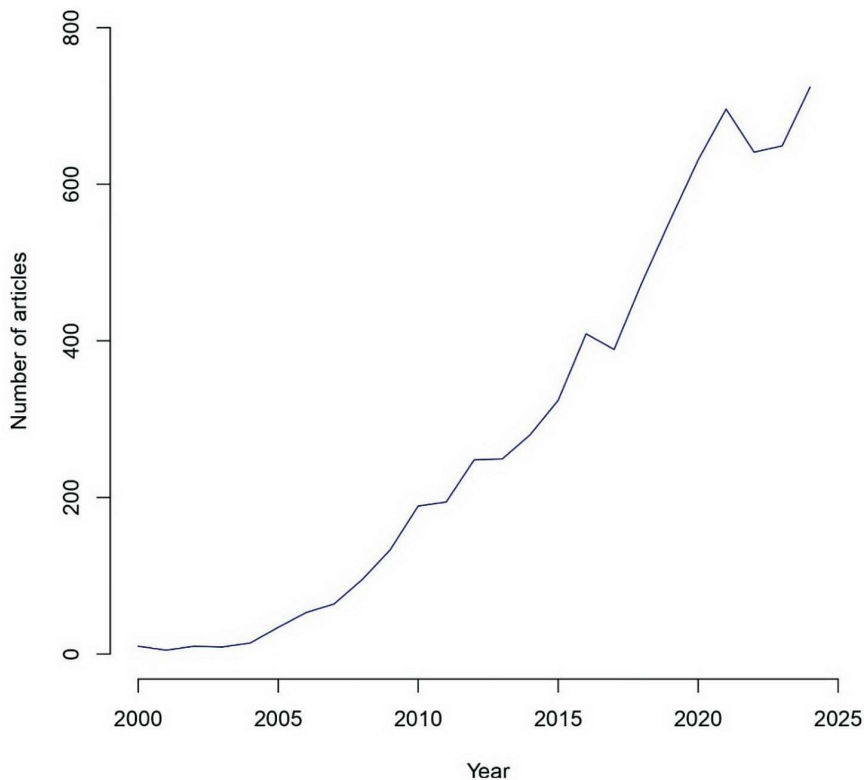
Source: Own preparation.

Figure 1. Number of articles per affiliation



Source: Own preparation.

Figure 2. Number of articles per year



Source: Own preparation.

Figure 3. Word clouds

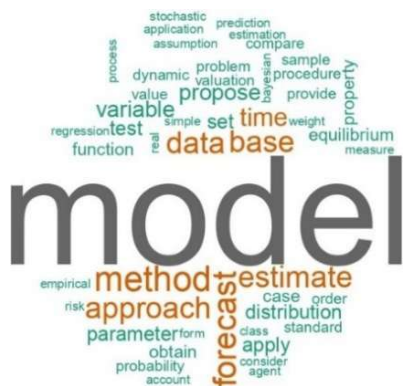


(1) Entrepreneurship



(2) Financial markets

Figure 3 cont. Word clouds



(3) Econometric modelling and forecasting



(4) Innovations



(5) Regional studies



(6) Economic growth



(7) Foreign trade



(8) Labour market

Source: Own preparation.

(9) Banking and finance

(10) Income and inequality

(11) Consumer behaviour

(12) Monetary and fiscal policy

(13) European Union studies

(14) Management

Source: Own preparation.



Figure 3 cont. Word clouds



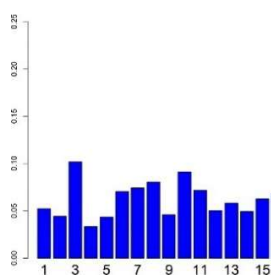
(15) Public policy and sustainability



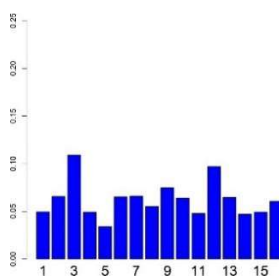
(16) Institutional economics

Source: Own preparation.

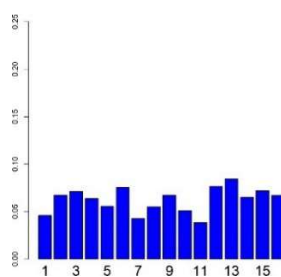
Figure 4. Topic prevalence: mean weights



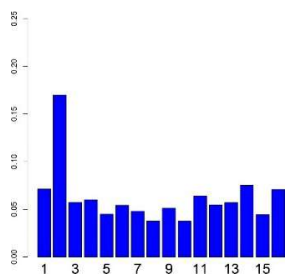
(1) University of Warsaw



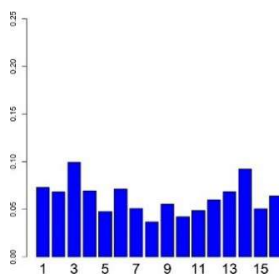
(2) Warsaw School of Economics



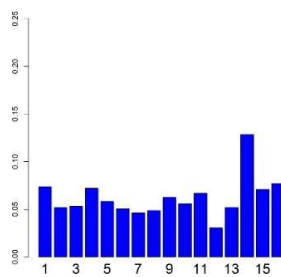
(3) University of Lodz



(4) Poznan University of Economics and Business



(5) Cracow University of Economics

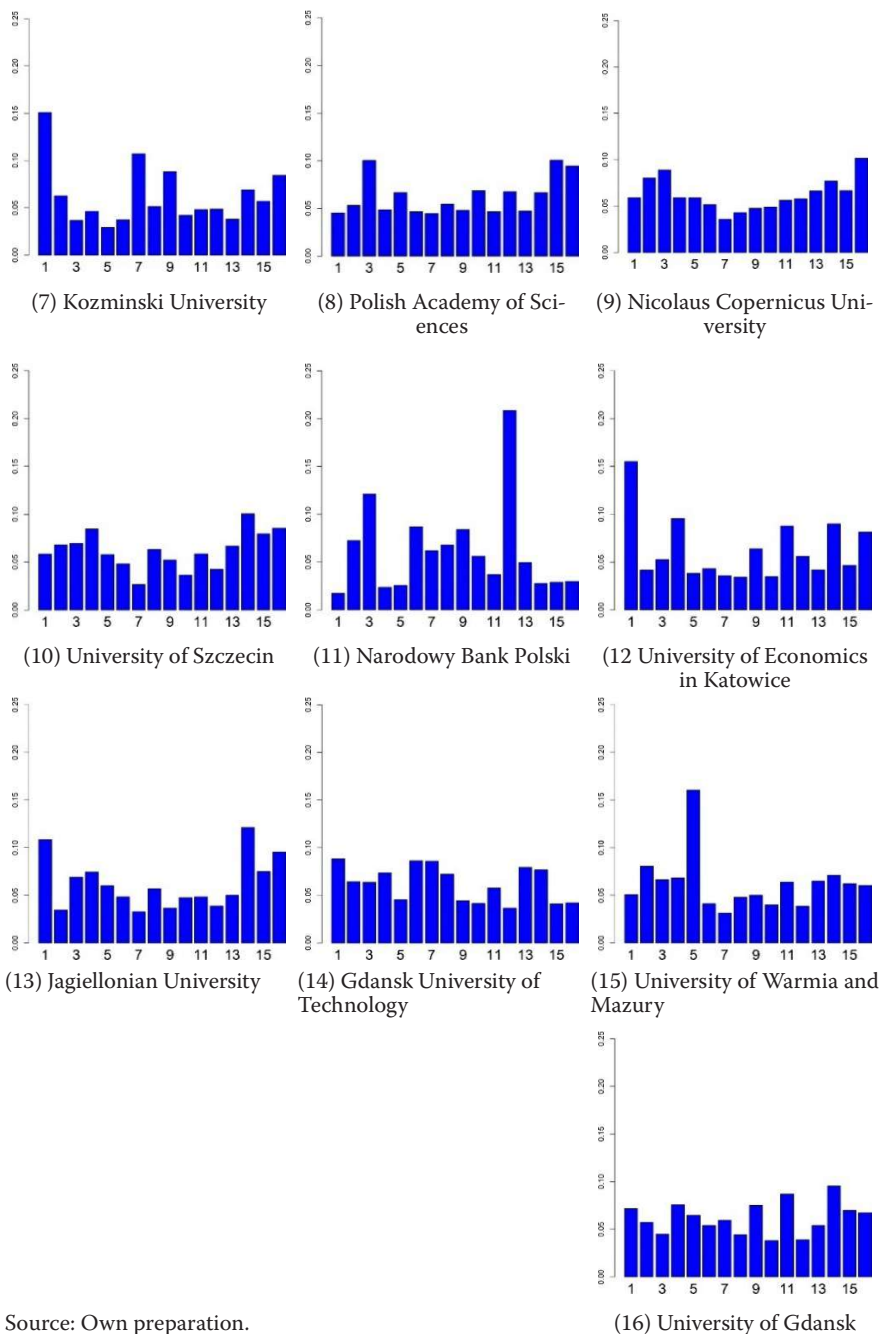


(6) Wroclaw University of Economics

Source: Own preparation.



Figure 4 cont. Topic prevalence: mean weights



Source: Own preparation.



Table 4. Top topics for research institutions: dominant topic

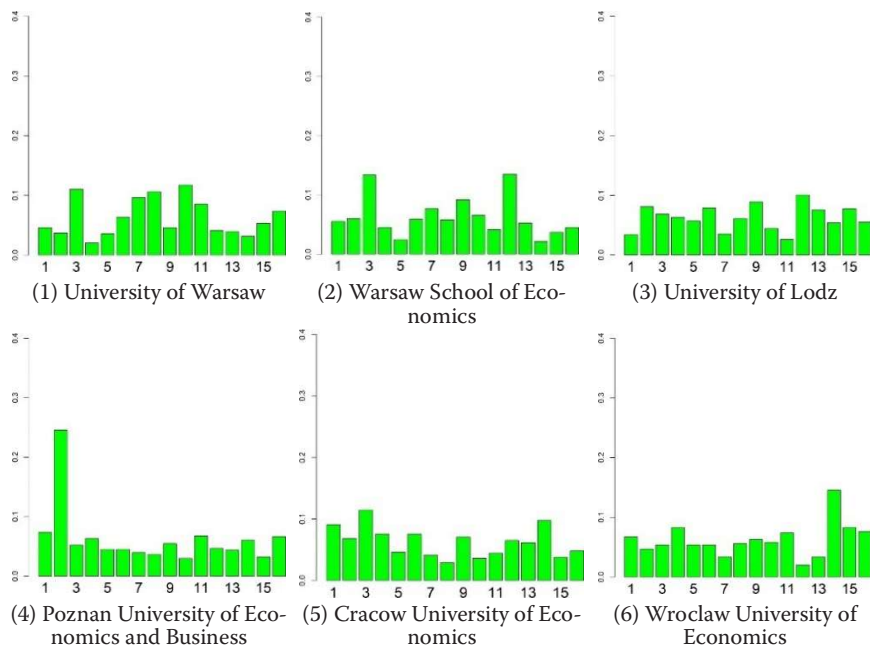
Research Institution	Top topics	Topic weights
University of Warsaw	10 – Income & inequality	0.1172
	3 – Econometric modelling & forecasting	0.1099
	8 – Labour market	0.1062
Warsaw School of Economics	12 – Monetary & fiscal policy	0.1347
	3 – Econometric modelling & forecasting	0.1338
	9 – Banking & finance	0.0917
University of Lodz	12 – Monetary & fiscal policy	0.1004
	9 – Banking & finance	0.0890
	2 – Financial markets	0.0807
Poznan University of Economics and Business	2 – Financial markets	0.2453
	1 – Entrepreneurship	0.0735
	11 – Consumer behaviour	0.0675
Cracow University of Economics	3 – Econometric modelling & forecasting	0.1145
	14 – Management	0.0974
	1 – Entrepreneurship	0.0906
Wroclaw University of Economics and Business	14 – Management	0.1454
	4 – Innovations	0.0828
	15 – Public policy & sustainability	0.0828
Kozminski University	1 – Entrepreneurship	0.2182
	9 – Banking & finance	0.1432
	7 – Foreign trade	0.1386
Polish Academy of Sciences	15 – Public policy & sustainability	0.1198
	16 – Institutional economics	0.1198
	3 – Econometric modelling & forecasting	0.1048
Nicolaus Copernicus University	16 – Institutional economics	0.1084
	2 – Financial markets	0.1024
	3 – Econometric modelling & forecasting	0.0934
University of Szczecin	14 – Management	0.1029
	4 – Innovations	0.0997
	8 – Labour market	0.0900
Narodowy Bank Polski	12 – Monetary & fiscal policy	0.3111
	3 – Econometric modelling & forecasting	0.1556
	9 – Banking & finance	0.1074
University of Economics in Katowice	1 – Entrepreneurship	0.2000



Research Institution	Top topics	Topic weights
	11 – Consumer behaviour	0.1185
	4 – Innovations	0.0963
Jagiellonian University	1 – Entrepreneurship	0.1504
	14 – Management	0.1463
	15 – Public policy & sustainability	0.0894
Gdansk University of Technology	7 – Foreign trade	0.1106
	1 – Entrepreneurship	0.0979
	8 – Labour market	0.0894
University of Warmia and Mazury	5 – Regional studies	0.2723
	2 – Financial markets	0.1149
	4 – Innovations	0.0681
University of Gdansk	14 – Management	0.1171
	9 – Banking & finance	0.1024
	11 – Consumer behaviour	0.0976

Source: Own preparation

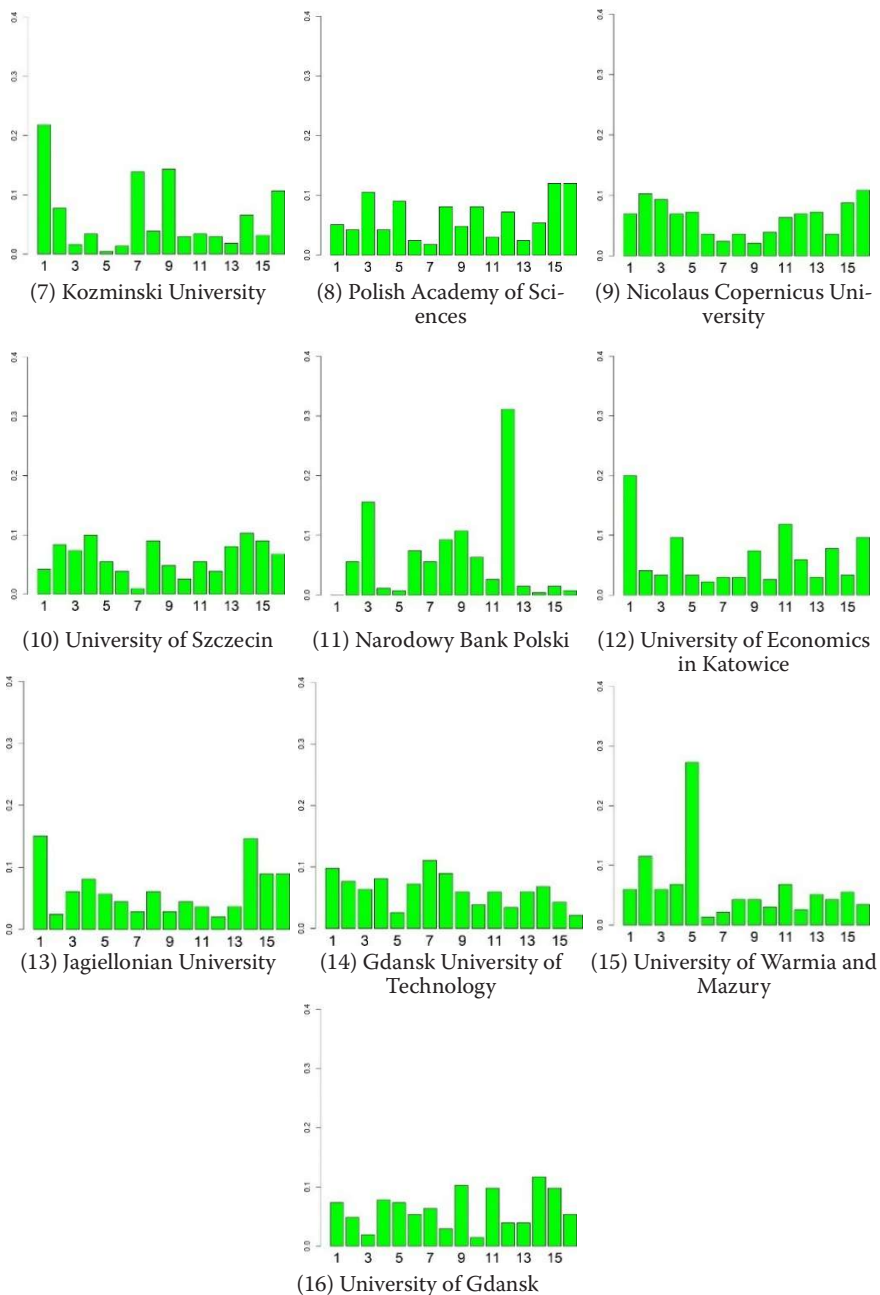
Figure 5. Topic prevalence: dominant topic



Source: Own preparation



Figure 5 cont. Topic prevalence: dominant topic



Source: Own preparation.

