Comparative Economic Research. Central and Eastern Europe
Volume 27, Number 2, 2024
https://doi.org/10.18778/1508-2008.27.10

Comparing Links between Topic Trends and Economic Indicators in the German and Polish Academic Literature

Victor Bystrov https://orcid.org/0000-0003-0980-2790
Assistant Professor, University of Lodz, Lodz, Poland, e-mail: victor.bystrov@uni.lodz.pl

Viktoriia Naboka-Krell https://orcid.org/0000-0003-0690-2737
Justus Liebig University of Giessen, Giessen, Germany, e-mail: viktoriia.naboka@wirtschaft.uni-giessen.de

Anna Staszewska-Bystrova https://orcid.org/0000-0002-3941-4986
Associate Professor, University of Lodz, Lodz, Poland, e-mail: anna.bystrova@uni.lodz.pl

Peter Winker https://orcid.org/0000-0003-3412-4207
Professor Justus Liebig University of Giessen, Giessen, Germany, e-mail: peter.winker@wirtschaft.uni-giessen.de

Abstract

The popularity of econometric analyses that include variables obtained from text mining is growing rapidly. A frequently applied approach is to identify topics from large corpora, which makes it possible to determine trends that reflect the changing relevance of topics over time. We address the question of whether such topic trends are linked to quantitative economic indicators typically used for analysing the objects described by a topic. The analysis is based on academic economic articles from Poland and Germany from 1984 to 2020. There is a specific focus on whether relationships between topic trends and indicators are similar across national economies. The connection between topic trends and indicators is analysed using vector autoregressive models and Granger causality tests.

Keywords: topic modelling, text analysis, latent Dirichlet allocation, Granger causality, topic trends

JEL: C49

Received: 18.04.2023. Verified: 21.09.2023. Accepted: 4.04.2024

Introduction

The popularity of analyses that include variables obtained from texts in econometric models is growing rapidly. Identifying topics in large corpora is often done as it makes it possible to determine trends that reflect the changing relevance of topics over time. Such trends might provide information about economic developments, and they can subsequently be used as (additional) indicators in economic analyses. Therefore, we determine whether topic trends obtained from academic text corpora from Poland and Germany might be linked to quantitative economic indicators typically used for analysing the objects described by a topic. The textual data cover the period 1984–2020. Given the different developments in Poland and Germany during this period, we are particularly interested in whether the relationships between topic trends and indicators were similar across national economies. Consequently, we focus on topics which are common to both text corpora.

The underlying textual data concern empirical economics and applied statistics (the German text collection) and econometric modelling in general (the Polish text collection). Given these specific characteristics of the two corpora, the uncovered topics are either associated with economic phenomena or the methods used to analyse them. The relative interest over time in both categories of topics can be potentially linked to real economic processes as certain developments can directly motivate the discussion in the literature and be reflected in the changing popularity of particular methods or frameworks of doing the analysis. Since the first type of link might be more pronounced, and as themes related to methods or theory found for both collections of texts were quite general, we focus on topics with macroeconomic content in this contribution.

The goal of the analysis is twofold. Firstly, we provide further evidence that the results of topic modelling, i.e. uncovered latent topics and topic weight time series, can be used in the next step to study the links between real developments and the popularity of topics in a given text corpus. This type of result provides new insights for describing a text collection. Secondly, we compare the links between interest in topics common to Polish and German text corpora and developments in the national and global economies. Thereby, given the differences between the corpora and the economies considered, it is not obvious how many common topics can be identified. The paper contributes to the literature on the links between textual data and real indicators by considering topics which were discussed in texts published in two countries and by providing results of causality tests for both countries.

Establishing links between textual data and quantitative economic indicators has been the subject of analysis in the past, with fast-growing interest during the last few years. Thereby, one may differentiate between keyword and sentiment-based methods and approaches that focus on identifying topics. Examples of the first group are the uncertainty

index proposed by Baker, Bloom, and Davis (2016) and, more recently, the fiscal sentiment indicator discussed by Latifi et al. (2023). Our contribution belongs to the second group, which also includes papers on using topic modelling in the context of innovation activities by Venugopalan and Rai (2015) and Bergeaud, Potiron, and Raimbault (2017). They analysed the classification of patents based on patent applications, though without an explicit focus on the time dimension. Furthermore, Lenz and Winker (2020) extracted innovation-related topics from news-ticker data and generated a time series of topic weights linked to fields on innovations. With a similar approach, Savin, Ott, and Konop (2022) provide an analysis of diffusion for different types of service robots.

Papers more closely related to our research compare times series of topic weights with quantitative economic indicators. Lüdering and Winker (2016) studied the dynamic links between the relevance of topics in academic publications and the development of corresponding macroeconomic indicators between 1949 and 2010 for Germany. Additionally, Hansen, McMahon, and Prat (2017) analysed the impact of increased transparency on monetary policy using publications of the Federal Open Market Committee. Dybowski and Adämmer (2018) used a topic model to analyse fiscal policy in the US, while Huang et al. (2018) and Larsen and Thorsrud (2022) considered the dynamic links between topic importance and financial market outcomes. Thorsrud (2020) and Ellingsen, Larsen and Thorsrud (2022) used related approaches to exploit textual data from newspapers to improve GDP forecasts.

The structure of the article is as follows. Section 2 presents the German and Polish text corpora, topics found using latent Dirichlet allocation, a topic-matching procedure, and real economic indicators. Section 3 describes time series methods used to analyse the relations between topic trends and economic indicators. The results on the links of topics and real economic developments are provided in Section 4. A summary of the central findings and suggestions for further work are given in Section 5.

Textual and economic data

The analysis is based on two different types of data: textual data, in the form of academic research articles, and standard economic variables. While the textual data must be transformed into quantitative indicators by means of topic modelling, the challenge regarding the economic indicators involves selecting appropriate series that correspond to the topics identified based on the textual data. The two types of data are described in the following two subsections.

Transforming text to time series

The text corpora comprise research papers in the field of economics and econometrics published from 1984 to 2020 in Germany and Poland. For Germany, we use all original articles from the Journal of Economics and Statistics (JES)[1], which were published mainly in German during the first half of the sample, then with an increasing share in English, reaching almost 100% towards the end of the sample. For Poland, we combine contributions to the proceedings of the Macromodels International Conference (MM) until 2011 and the Central European Journal of Economic Modelling and Econometrics (CEJEME) after 2009. Detailed descriptions of both text corpora can be found in Bystrov et al. (2022).

While the scope of both academic outlets is international, we observe that contributions discussing issues related to the German and Polish economies, respectively, constitute considerable shares of articles over the sample period. This is the rationale for contrasting the results from topic modelling with national as well as global economic indicators. Furthermore, given the low frequency of publications (six issues per year for JES, one for MM and four for CEJEME), the following analysis will be done at an annual level. Thereby, we also alleviate the problem of publication lags inherent both in conference contributions and even more so in journal publications. We will return to this issue in the discussion of the Granger causality analysis in Section 4.

The first step in transforming the text corpora into quantitative indicators involves applying a latent Dirichlet allocation (LDA) model to both corpora (Blei, Ng, and Jordan 2003). For details on pre-processing, parameter choice, and the handling of the multilingual German corpus, see again Bystrov et al. (2022). The LDA estimation results provide topics discussed in the text corpora and their relative weights, which change over time. These weights correspond to the topic time series, which are considered in the subsequent analysis.

The model selection step for the LDA indicated 37 and 60 topics for the Polish and German corpora, respectively. The larger number of different topics for the German dataset is plausible, given the broader scope of JES compared to MM and CEJEME. In further analysis, we focus on topics and their corresponding weight time series, which are relevant in both corpora. Common themes were identified using the topic-matching method proposed by Bystrov et al. (2022), which is based on a comparison of topic distributions over a joint vocabulary. Topic resemblance is evaluated using the cosine similarity measure, and the matching is done by finding the nearest neighbour of a topic estimated for one corpus in the other corpus. The pairs which are considered reasonable matches are selected based on a cut-off value for the cosine similarity measure, i.e.,

1 For a previous application of topic modelling to articles published in this journal see Lüdering and Winker (2016).

only matches with a high enough cosine similarity are kept. For the present application, this threshold was set to 0.265, resulting in 24 topics appearing in both corpora. The labels of the topics were assigned based on the inspection of the corresponding word clouds available in Bystrov et al. (2022) and the documents with the highest weights for the particular topic.

For further analysis, we restrict the set of topics to those that correspond to applied economic research, which might be more closely related to specific economic developments than the dynamics of topics related to purely methodological or theoretical research in economics, econometrics and statistics, such as forecasting, simulation methods or welfare economics. Furthermore, our focus is on the national economy level. Therefore, we chose topics that describe macroeconomic relationships rather than processes at the micro level, such as firm growth or household income. Appendix A.1 provides information on the topics which were rejected. Table 1 below lists the remaining topics, the corresponding cosine similarity values, and the economic indicators chosen as described in the following subsection.

Table 1. Summary of common topics and selected economic indicators

Topic label	Proposed indicator	Cosine similarity
International economics	Trade share	0.86 146
Banking and credit	Credit-to-GDP ratio	0.85 116
Business cycle	Output gap	0.83 445
Capital and growth	Growth rate of investment	0.80 610
Labour market	Unemployment rate	0.72 227
Crude oil market	Oil market shocks	0.70 459
Monetary policy	Policy rate	0.69 914
Stock market	Stock market return, stock price volatility	0.63 251
Foreign trade	Net export share	0.49 535
Energy sector	Total primary energy production	0.48 663

Source: authors' own elaboration.

Selecting economic indicators

The selection of economic indicators related to the topics shown in Table 1 is based on the interpretation of these topics and considering related articles with the highest weights. Furthermore, this selection must consider data availability for both countries over the sample period. The specific sources and available samples of these economic time series are summarised in Table 3 in Appendix A.2.

Given the structure of textual resources, all data are annual. This frequency of observations is appropriate for considering links between real developments and their discussion in journal and conference articles, as publication lags or time for researching a new topic might be considerable.

While for most of the topics a suitable observable economic indicator is assigned for both the German and the Polish economies, the setting is different for the topic that refers to the crude oil market. In this case, a suitable indicator would be one of the shocks to the international oil market. Given that such an indicator cannot be observed directly, it must be derived first using an auxiliary model. To this end, we use the method proposed by Kilian (2009). Then, given that global oil market shocks are considered, the same time series of shocks is used in the analysis of the Polish and German data. A more detailed discussion of the variables listed in Table 1 follows in Section 4, together with the results of Granger causality testing.

Methods

The quantitative analysis is based on vector autoregressive models (VARs), which are a natural choice if the aim is to test for (Granger) causality when the direction of causality is not known a priori. We use bivariate VAR models, where one of the variables is the weight of a topic aggregated over all documents for each period. We label these series topic weight series. The second variable is the economic indicator linked to the specific model. A two-dimensional VAR(p) is given by the following formula (see Kilian and Lütkepohl 2017):

$$y_t = \nu + A_1 y_{t-1} + A_2 y_{t-2} + \ldots + A_p y_{t-p} + B d_t + u_t, \tag{1}$$

where $y_t = (topic_t, ind_t)'$ is the vector of topic weight and economic indicator in period t. The parameter matrices $A_i (i = 1, \ldots, p)$ have dimensions 2×2, ν is a two-dimensional intercept vector, d_t includes all remaining necessary deterministic terms like dummy variables or trends with the corresponding parameters gathered in matrix B, and u_t is a two-dimensional vector of error terms.

Prior to estimation, variables included in the model are differenced if the results of the augmented Dickey-Fuller (ADF) test indicate the presence of a unit root. Furthermore, to consider the trending and nonlinear behaviour of topic weights, which can be observed in some cases, vector d includes the deterministic trend t and the second power of this variable in selected models. The lag order of the VAR, $\hat{p}$, is chosen using the Akaike information criterion (AIC), applying a maximal lag length of 4 years.

The models are tested for autocorrelation and autoregressive conditional heteroskedasticity (ARCH) effects of the error terms. Information on data transformations, the use of deterministic variables (apart from the intercept), and the selected lag order of the VAR for all the models are provided in Tables 5 and 6 in Appendix A.2.

In the last step, to check for dynamic links between economic indicators and topic trends, Granger causality tests are performed. The hypothesis of instantaneous causality is also tested. If autocorrelation or ARCH errors are detected, these tests are based on the heteroskedasticity- and autocorrelation-consistent (HAC) estimator of the variance matrix for the ordinary least squares (OLS) (see Table 6 in Appendix A.2 for the type of estimator used). The significance level for Granger causality tests is set to $\alpha = 0.1$. The outcomes of the tests may be to detect no causality, causality in one direction, or causality in both directions. If the topic weights prove to be Granger causal for the real economic indicators, this implies that changes in discussion intensity of the topic in the academic literature preceded relevant economic developments. The opposite outcome indicates that relevant developments in parts of the economy led to a more intensive academic discussion of these aspects afterwards. If Granger causality is found for both directions, both channels are relevant, i.e., specific economic developments are accompanied by a change in the relevance of the topic in science, both ex-ante and ex-post. While Granger causality focuses on the dynamic interlink between variables, it does not cover mutual influences taking place within one period, i.e., one year. Such contemporaneous effects are reflected by a significant correlation of the error terms of the VAR model across equations and are often labelled instantaneous causality. Given the low frequency of our data, instantaneous causality might comprise links that would be measured as Granger causality at a higher frequency. This effect could be quite pronounced as academic publications and conference proceedings were published with substantial publication lags for the period considered in our empirical application. Therefore, we also report the results of the tests of instantaneous causality.

Results

In this section results of the Granger causality analysis are presented for the selected topics common to the JES and MM/CEJEME text corpora as described in Section 2. Figures 1 and 2 present the distribution of the relative topic importance for both corpora, that is measured by computing the mean of topic weights over the sample period. The topics are ordered in descending order according to their relative importance for the respective corpora. The bars in orange represent "matched" topics, meaning that they are common for both countries, as identified by Bystrov et al. (2022) (see subsection 2.1). Correspondingly, the bars in blue represent "unmatched" topics, meaning that these topics are specific to a single corpus. Dashed orange bars highlight topics

selected for further analysis within the VAR models together with corresponding economic indicators (see Table 1 in Section 2).

The outcomes of the Granger causality testing are summarised in Table 2. These results are discussed in more detail for each topic separately in the following subsections. Word clouds for all the topics are shown in Figure 3.

Capital and growth

The topic of capital and growth was relatively important in both the MM/CEJEME and JES corpora. Average weights amounted to 0.0472 and 0.0354, respectively, which meant that it was the 4th and the 10th most discussed theme in respective text collections. The time series of topic weights were modelled together with an economic indicator given by the growth rate of gross fixed capital formation. Statistical tests (see Table 2) indicated Granger causality from the economic indicator to topic weights for the Polish texts and two-way Granger causality for the German corpus.

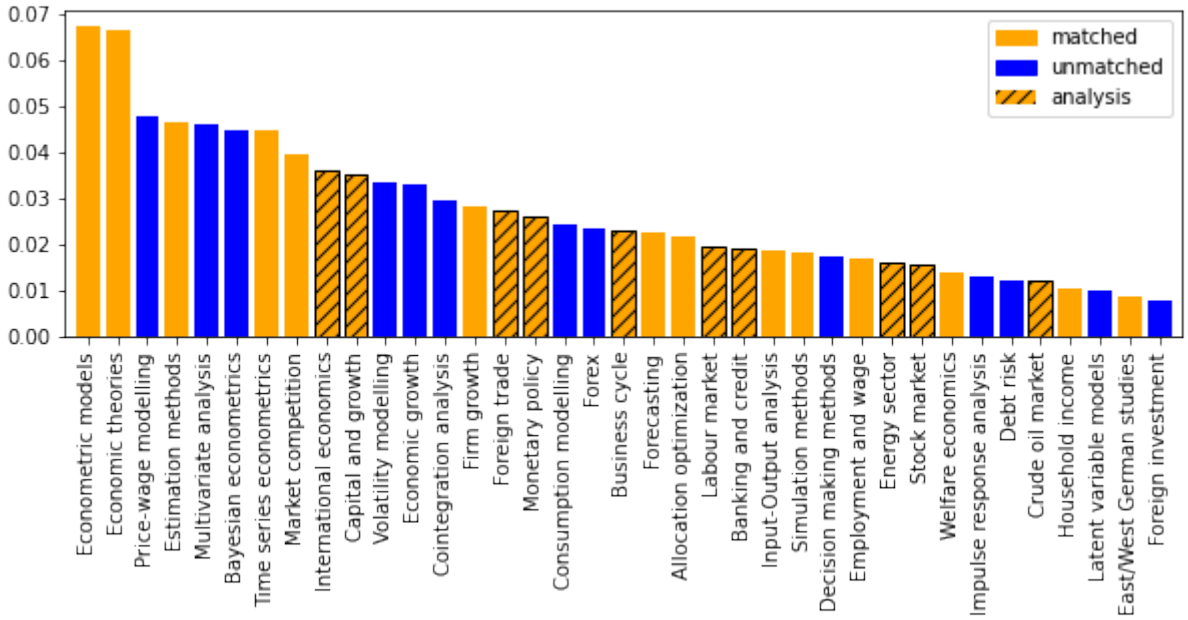

Figure 1. Relative topic importance for Poland

Source: authors' own elaboration.

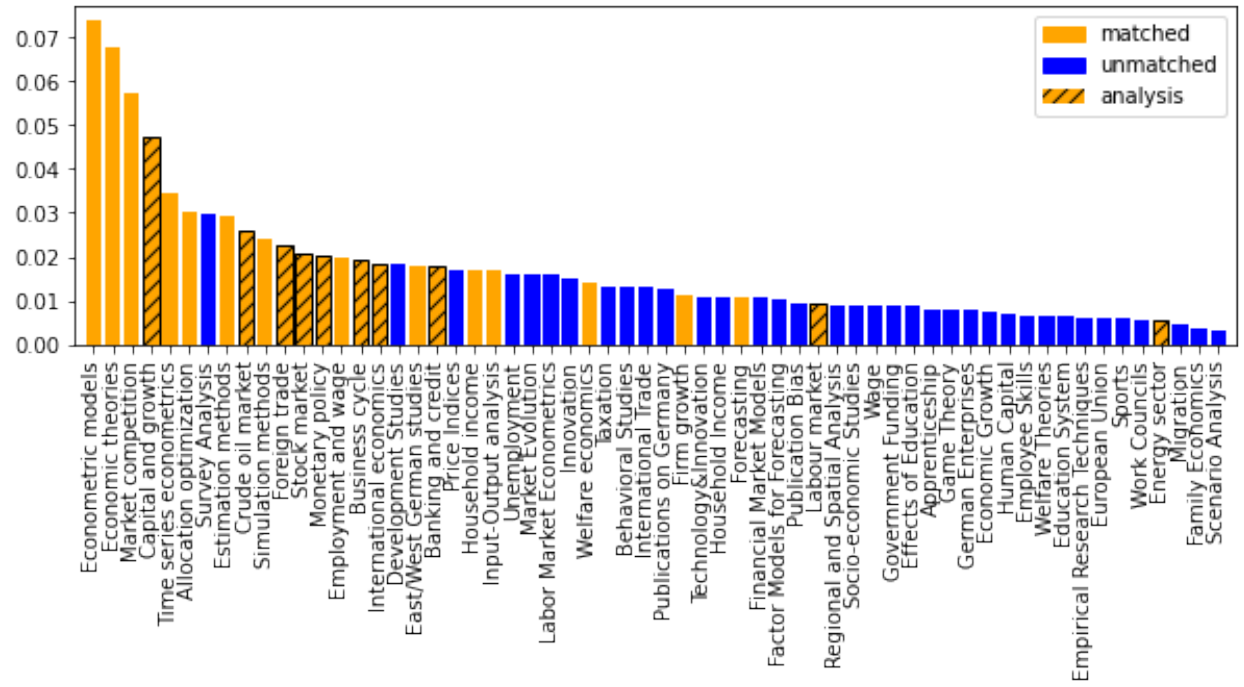

Figure 2. Relative topic importance for Germany

Source: authors' own elaboration.

Table 2. Granger causality test results

Topic label	Economic indicator	Country	Hypothesis:		
			The topic weight is not a Granger cause for the economic indicator	The economic indicator is not a Granger cause for the topic weight	No instantaneous causality
Capital and growth	Growth rate of gross fixed capital formation	Poland	1.5933 [0.2114]	**2.4496 [0.0763]**	0.7304 [0.4710]
		Germany	**2.3099 [0.0470]**	**4.3290 [0.0226]**	−0.5192 [0.6071]
International economics	Trade share	Poland	0.2718 [0.6069]	**3.0073 [0.0951]**	−0.7904 [0.4362]
		Germany	0.0314 [0.8605]	0.4400 [0.5122]	−1.2610 [0.2162]
Foreign trade	Share of net exports in GDP	Poland	**9.4156 [0.0005]**	**13.790 [0.0001]**	**−3.2370 [0.0034]**
		Germany	**7.3949 [0.0007]**	1.9984 [0.1317]	−0.5633 [0.5774]
Monetary policy	Policy rate	Poland	1.2599 [0.2719]	**4.2389 [0.0497]**	−0.3062 [0.7616]
		Germany	0.3366 [0.5658]	**3.7268 [0.0622]**	0.6340 [0.5303]
Business cycle	Output gap	Poland	1.6602 [0.2094]	0.1777 [0.6770]	−0.5225 [0.6054]
		Germany	2.1504 [0.1192]	1.1875 [0.3346]	−0.5355 [0.5960]
Crude oil market	Oil supply shocks	Poland	0.0835 [0.7746]	0.2850 [0.5972]	**3.0807 [0.0041]**
		Germany	1.3461 [0.2543]	1.3274 [0.2576]	0.4892 [0.6278]

Topic label	Economic indicator	Country	Hypothesis:		
			The topic weight is not a Granger cause for the economic indicator	The economic indicator is not a Granger cause for the topic weight	No instantaneous causality
Crude oil market	Aggregate demand shocks	Poland	1.4327 [0.2404]	0.0710 [0.7916]	– 1.0949 [0.2813]
		Germany	1.9530 [0.1716]	**3.3648 [0.0756]**	1.3285 [0.1929]
Crude oil market	Oil specific-demand shocks	Poland	0.3835 [0.5402]	**3.7426 [0.0622]**	– 0.8218 [0.4169]
		Germany	0.0748 [0.7862]	2.0362 [0.1630]	0.4259 [0.6728]
Banking and credit	Credit-to GDP ratio	Poland	2.1402 [0.1565]	1.0865 [0.3076]	– 1.0753 [0.2925]
		Germany	0.0670 [0.7975]	**7.0072 [0.0125]**	– 0.6651 [0.5106]
Stock market	Stock market return	Poland	0.3974 [0.5349]	0.2459 [0.6249]	– 0.4490 [0.6576]
		Germany	**5.3744 [0.0280]**	0.3702 [0.5478]	1.5390 [0.1346]
Stock market	Stock price volatility	Poland	**8.5297 [0.0022]**	**15.8720 [0.0002]**	– 0.1086 [0.9146]
		Germany	1.0500 [0.3140]	1.9445 [0.1738]	– 0.3685 [0.7151]
Labour market	Unemployment rate	Poland	0.4140 [0.5256]	0.1316 [0.7197]	**4.8774 [0.0000]**
		Germany	2.3394 [0.1153]	**3.6196 [0.0405]**	– 0.9967 [0.3264]
Energy sector	Primary energy production	Poland	**3.1800 [0.0862]**	0.0002 [0.9896]	0.2342 [0.8166]
		Germany	1.0585 [0.4107]	**4.1260 [0.0189]**	**– 4.1868 [0.0003]**

Source: authors' own elaboration.

International economics

The topic of international economics was relatively important for both text corpora. In the Polish text collection, it was the 9th most popular theme (with an average weight of 0.0363), and in the German corpus, it was 16th (with an average weight of 0.0185). As can be seen in the word clouds associated with this theme shown in Figure 3, even though these topics were paired based on the topic matching algorithm, they seem to concern slightly different matters as the German topic is more related to cross-country comparisons and international spillovers, while the Polish topic puts more emphasis on international trade. This different focus complicates the selection of an economic indicator, which was eventually specified as trade share, defined as the sum of the nominal value of imports and exports expressed as a per cent of nominal GDP.

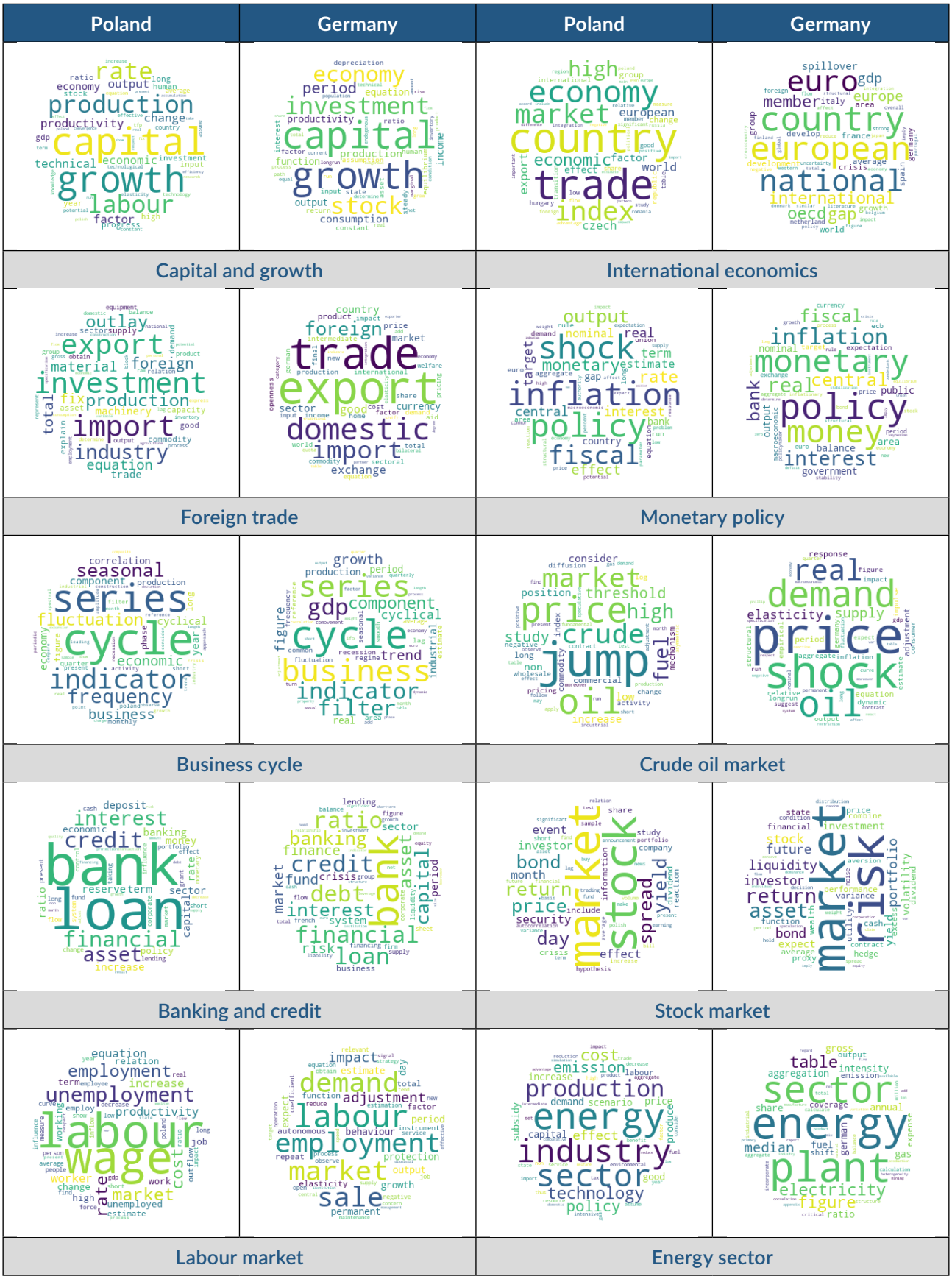

Figure 3. Word clouds

Source: authors' own elaboration.

According to the results provided in Table 2, no Granger causality between topic weights and the indicator was found for Germany, while shocks in international trade Granger caused the topic weights for Poland. For both countries, no significant instantaneous link was detected.

Foreign trade

With average weights of 0.0276 and 0.0224, the topic of foreign trade was the 15th and the 11th most discussed topic in the Polish and German text corpora, respectively. The economic indicator used in the VAR associated with foreign trade was the share of net exports in GDP. This variable was selected as the primary indicator of developments in foreign trade.

The outcomes of the Granger causality tests presented in Table 2 indicate two-way causality and instantaneous causality found for Poland and one-way causality for Germany. In the latter case, the changes in popularity of the subject in the academic literature preceded real movements in the German share of net exports in GDP.

Monetary policy

The topic labelled monetary policy was quite popular in both text corpora – it had the 16th and 13th highest mean weights in the Polish and German collections, respectively. A natural economic indicator for this topic is the monetary policy rate. Granger causality test results based on VAR that incorporated both policy rate and topic trend (see Table 2) indicate causality from the indicator to topic weights for both countries. This implies that past changes in the monetary policy rate add predictive power to the dynamic model to explain the development of topic weights over time.

Business cycle

The topic of business cycle was the 19th and 15th most popular in the Polish and German collections, respectively (the mean weights were 0.0232 and 0.0194). Bivariate VAR models were constructed for topic weights and national output gap indicators. However, no Granger causality was found for either corpus, nor was there an indication of instantaneous effects. Given the recurrence of business cycles, it might not come as a surprise that the academic literature on the topic does not lead or follow actual business cycles. Thus, the result of no Granger causality is sensible.

Crude oil market

Word clouds associated with the topic of the crude oil market (see Figure 3) show a clear focus on oil price shocks. This theme was relatively more important in the German corpus, where it was the 9th most frequently discussed topic (with an average weight of 0.0259). In the Polish text collection, it was only the 33rd most important of the 37 themes identified (with an average weight of 0.0121).

Given the word clouds, it is intuitive that the corresponding topic weights might be related to shocks in the global crude oil market rather than to an oil price series. Therefore, for this specific topic, the relevant economic indicator is not directly observable, but must be derived in a first step. To construct a series of shocks, we used the method described by Kilian (2009). It is based on a trivariate structural VAR, including log differences of world crude oil production ($\Delta prod$), an index of global real economic activity (*rea*), and the real price of oil (*rpo*) (see Table 4 for the description of the data). As recommended by Kilian (2009), the model is estimated using monthly data. The model has the following form:

$$A_0 y_t = \nu + \sum_{i=1}^{24} A_i y_{t-i} + u_t,\ A_0^{-1} = \begin{bmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

where $y_t = (\Delta prod_t, rea_t, rpo_t)'$. The three structural shocks that Kilian (2009) labelled oil supply shocks, aggregate demand shocks, and oil-specific demand shocks are given by

$$e_t = A_0^{-1} \begin{bmatrix} u_t^{oil\ supply\ shock} \\ u_t^{aggregate\ demand\ shock} \\ u_t^{oil\ specific-demand\ shock} \end{bmatrix},$$

and represent innovations in global oil production, global demand for industrial commodities, and precautionary demand for oil, respectively.

These estimated series of structural shocks are aggregated to annual frequency as used in the present study. Unlike in other VAR models, where topic weights from the German and Polish text collections were modelled jointly with economic indicators from Germany and from Poland, respectively, in the models used to study the topic of the crude oil market, the same series of global shocks are used for both countries.

As shown in Table 2, for Poland, oil-specific demand shocks were found to be a Granger cause for the topic weights, and instantaneous causality was found for oil supply

shocks. This implies that future academic discussion on the role of the crude oil market is stimulated by demand shocks, while supply shocks might be discussed faster or – given the lag between academic research and publications – even with some lead. For Germany, we expected a similar finding in particular due to the high interest in economic science devoted to the crude oil market after the supply shocks in the 1970s and early 1980s. However, since our sample starts in 1984, these events are not part of the sample period. This might explain why no Granger causality was found for German data for the models that incorporate oil supply shocks and oil-specific demand shocks. In contrast, a surge in interest in the topic between 2002 and 2005 correlates with large shocks to the index of global demand for industrial commodities (the Kilian index, see Kilian 2009). Shocks to the global demand for industrial commodities can be transmitted (with a delay) to crude oil prices and eventually to the discussion of oil prices in the JES, which might explain the finding that aggregate demand shocks were found to be a Granger cause for crude oil market topic weights.

Banking and credit

The topic of banking and credit was the 23rd most covered theme in the Polish corpus. In the German collection, it ranked 19th. In the Granger causality analysis, topic weights were contrasted with an economic indicator for the importance of the banking sector given by the credit-to-GDP ratio. This variable proved to be Granger causal for topic weights for the German corpus, i.e., a shock to the credit-to-GDP ratio preceded changes in the importance of the topic in the academic literature. However, no Granger causality was found for the Polish text collection, nor was there an indication of contemporaneous effects.

Stock market

The topic related to the stock market attracted more attention in the German corpus, where it came 12th with an average weight of 0.0205. In the Polish text collection, it was 29th and had a mean weight of 0.0156. This discrepancy in the relative interest was expected as the Warsaw Stock Exchange did not operate during the first part of our sample period (it became active in 1991; our sample comprises data from 1995 to 2020 for Poland).

For this topic, two economic indicators were considered alternatively, as it was difficult to select the best-suited one a priori. The two indicators were the stock market return, computed as the growth rate of the annual average national stock market index, and stock price volatility, calculated as the average of the 360-day volatility of the national stock market index. The two indices considered were Deutscher Aktienindex

(DAX) and Warszawski Indeks Giełdowy (WIG), for German and Polish corpora, respectively.

The results from Table 2 indicate two-way Granger causality between stock price volatility and the topic weights for the Polish corpus. For the German text collection, topic weights were a Granger cause for the stock market returns series.

Labour market

The most important words for the labour market (see Figure 3) suggest that the topic in the German corpus has a broad focus on the labour market and its functioning. In contrast, the Polish corpus focuses more on wage and (un-)employment. The weights of this topic were 0.0198 and 0.0092 for the Polish and German corpora, respectively. For both text collections, the topic belonged to the bottom half of the most popular themes (22nd place out of 37 for Poland and 39th out of 60 for Germany). Topic weights were modelled jointly with the unemployment rate. The outcomes of Granger causality tests were as follows: causality from indicator to topic weights was found for the German corpus, and instantaneous causality was detected for the Polish corpus.

Energy sector

The energy sector was not a very popular topic in either corpus. With an average weight of 0.0159, it was 28th on the list of most frequently discussed topics for Poland. For the German corpus, it was the third least important theme (with a mean weight of only 0.0053, it occupied 57th place). The economic indicator selected for causality analysis was primary energy production.

The results from Table 2 show significant links between topic weights and the economic indicator for both corpora. The direction of causality is, however, different: for the German corpus, real developments precede and are simultaneous with the changes in topic weights, while there is causality from topic weights to the real indicator in the case of the Polish corpus.

Conclusions

We used text mining techniques to identify topics discussed in two academic text corpora on economics from Germany and Poland for the same period, 1984–2020. Our main focus was on the relative importance of these topics over time and how they corresponded to central economic indicators related to the topics. For one theme, which corresponded to shocks to the crude oil market, the indicator had to be derived from an auxiliary model.

In order to determine whether economic research led or followed real events, bivariate relationships between topic trends and economic variables were analysed using vector autoregressive models and Granger and instantaneous causality tests. We considered ten distinct topics which appeared in both text corpora. This allowed us to study whether Granger causality between topic trends and the corresponding indicators was frequent and also which direction of causality prevailed. Furthermore, parallel analysis for German and Polish textual data and economic indicators made it possible to compare results for these two countries.

The analysis indicated significant links between academic literature and real developments for at least one country for all but one pair of topic and the corresponding economic indicator. This exception was given by the topic of business cycle and the output gap. In this case, academic analysis might be focused more on the existence and stability of the recurring cycles than on each particular cycle, which might explain the lack of a significant link.

In general, the reason for not detecting causality between topics and indicators might also be due to the relatively small sample size, the imperfect matching of topics from the German and Polish corpora and, consequently, problems with selecting an appropriate economic indicator that fits the topics for both countries. Therefore, the substantial number of significant links actually found indicates a strong focus of researchers publishing in the two corpora on empirical evidence. However, the results of the direction of Granger causality were mixed. Examples of instantaneous causality, both kinds of unidirectional causality, and two-way causality were found. Overall, it can be concluded that economic indicators led topic trends more often than the other way round. Thus, economists tend to follow real developments in their analysis rather than predict them.

While the frequency of significant links between economic variables and topic popularity for the Polish and German corpora were comparable, causality patterns for specific topic-indicator pairs were often not the same. Similarities included, for example, significant reactions by researchers of the topic of monetary policy to changes in the policy rate and Granger causality from the growth rate of gross fixed capital formation to the trend of the topic capital and growth. Interestingly, the discussion in economic science of foreign trade preceded changes in the share of net exports in GDP in both countries, i.e., globalisation was discussed by researchers prior to its realisation. Some differences in the results across countries, e.g., for the topics on the stock market or the energy sector, can be explained by the different economic settings in Poland and Germany during the years under investigation. For example, the stock exchange only emerged in Poland during the sample period, and energy production mixes were not comparable, which might be reflected in the focus and content of economic discussion of these topics.

There are several limitations of our analysis, which might be considered a first explorative study of the link between economic science and economic reality in a cross country comparison. Further research should address both methodological and content-related issues. On the methodological side, alternative methods for identifying (common) topics and quantifying their relevance over time might be considered. As far as content is concerned, it might be of interest to include text corpora from other countries and to extend the observation period. Further research should also focus on the link between topics and economic reality for specific topics, possibly using case studies for a better understanding of the driving forces behind the links found in our analysis.

Acknowledgements

Financial support from the German Research Foundation (DFG) (WI 2024/8–1) and the National Science Centre (NCN) (Beethoven Classic 3: UMO–2018/31/G/HS4/00869) for the project TEXTMOD is gratefully acknowledged. The project also benefited from cooperation within HiTEC Cost Action CA 21163.

References

Baker, S.R., Bloom, N., Davis, S.J. (2016), *Measuring Economic Policy Uncertainty*, "The Quarterly Journal of Economics", 131 (4), pp. 1593–1636, https://doi.org/10.1093/qje/qjw024

Bergeaud, A., Potiron, Y., Raimbault, J. (2017), *Classifying patents based on their semantic content*, "PLOS ONE", 12 (4), pp. 1–22, https://doi.org/10.1371/journal.pone.0176310

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

Bystrov, V., Naboka, V., Staszewska-Bystrova, A., Winker, P. (2022), *Cross-Corpora Comparisons of Topics and Topic Trends*, "Journal of Economics and Statistics", 242 (4), pp. 433–469, https://doi.org/10.1515/jbnst-2022-0024

Dybowski, T.P., Adämmer, P. (2018), *The economic effects of U.S. presidential tax communication: Evidence from a correlated topic model*, "European Journal of Political Economy", 55, pp. 511–525, https://doi.org/10.1016/j.ejpoleco.2018.05.001

Ellingsen, J., Larsen, V.H., Thorsrud, L.A. (2022), *News media versus FRED-MD for macroeconomic forecasting*, "Journal of Applied Econometrics", 37 (1), pp. 63–81, https://doi.org/10.1002/jae.2859

Hansen, S., McMahon, M., Prat, A. (2017), *Transparency and Deliberation Within the FOMC: A Computational Linguistics Approach*, "The Quarterly Journal of Economics", 133 (2), pp. 801–870, https://doi.org/10.1093/qje/qjx045

Huang, A.H., Lehavy, R., Zang, A.Y., Zheng, R. (2018), *Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach*, "Management Science", 64 (6), pp. 2833–2855, https://doi.org/10.1287/mnsc.2017.2751

Kilian, L. (2009), *Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market*, "American Economic Review", 99 (3), pp. 1053–1069, https://doi.org/10.1257/aer.99.3.1053

Kilian, L., Lütkepohl, H. (2017), *Structural Vector Autoregressive Analysis*, "Themes in Modern Econometrics", Cambridge University Press, Cambridge.

Larsen, V.H., Thorsrud, L.A. (2022), *Asset returns, news topics, and media effects*, "The Scandinavian Journal of Economics", 124 (3), pp. 838–868, https://doi.org/10.1111/sjoe.12469

Latifi, A., Naboka-Krell, V., Tillmann, P., Winker, P. (2023), *Fiscal policy in the Bundestag: Textual analysis and macroeconomic effects*, MAGKS papers on economics, Philipps-Universität Marburg, Faculty of Business Administration and Economics, Department of Economics, Marburg.

Lenz, D., Winker, P. (2020), *Measuring the diffusion of innovations with paragraph vector topic models*, "PLOS ONE", 15 (1), pp. 1–18, https://doi.org/10.1371/journal.pone.0226685

Lüdering, J., Winker, P. (2016), *Forward or Backward Looking? The Economic Discourse and the Observed Reality*, "Journal of Economics and Statistics", 236 (4), pp. 483–515, https://doi.org/10.1515/jbnst-2015-1026

Savin, I., Ott, I., Konop, C. (2022), *Tracing the evolution of service robotics: Insights from a topic modeling approach*, "Technological Forecasting & Social Change", 174, 121280, https://doi.org/10.1016/j.techfore.2021.121280

Thorsrud, L.A. (2020), *Words are the new numbers: A newsy coincident index of the business cycle*, "Journal of Business & Economic Statistics", 38 (2), pp. 393–409, https://doi.org/10.1080/07350015.2018.1506344

Venugopalan, S., Rai, V. (2015), *Topic based classification and pattern identification in patents*, "Technological Forecasting and Social Change", 94, pp. 236–250, https://doi.org/10.1016/j.techfore.2014.10.006

Porównanie zależności pomiędzy popularnością tematów artykułów naukowych i zmiennymi ekonomicznymi dla Polski i Niemiec

W ostatnim czasie obserwować można gwałtowny wzrost popularności metod analizy ekonomicznej wykorzystującej zmienne pozyskane z tekstów. Jednym z najczęściej stosowanych podejść jest modelowanie tematów, które pozwala na oszacowanie, jak waga poszczególnych tematów zmieniała się w czasie. Celem artykułu jest zbadanie, czy mierzona za pomocą wag popularność tematów była powiązana z wybranymi zmiennymi ekonomicznymi. W badaniu wykorzystano artykuły naukowe z obszaru ekonomii, opublikowane w Polsce i Niemczech w latach 1984–2020. Jednym z celów analizy było stwierdzenie, czy zależności pomiędzy popularnością wybranych tematów w Polsce i w Niemczech i powiązanymi z nimi wskaźnikami ekonomicznymi były podobne. Badanie przeprowadzono za pomocą modeli wektorowej autoregresji i testów przyczynowości Grangera.

Słowa kluczowe: modelowanie tematów, analiza tekstu, alokacja zmiennej ukrytej Dirichleta, przyczynowość w sensie Grangera, popularność tematów

Appendix

A. Data sources

A1. Selection of topics for further analysis

after the analysis of the topic word clouds, 10 out of 24 topics were selected for the analysis (see Table 1). Nine topics, labelled forecasting, time series econometrics, econometric models, estimation methods, economic theories, input-output analysis, simulation methods, allocation optimisation and welfare economics were identified as methodological or theoretical and thus disregarded from this study. Four further topics, corresponding to firm growth, market competition, household income, and employment and wage were also rejected as they seemed to correspond to the micro economic level, which might not be well reflected in the aggregated observed time series employed in this paper. Finally, the topic labelled East-West German studies was disregarded as it deals with studying the consequences of a specific event.

A2. Data sources and data transformations

Table 3. Description of data

Topic	Variable	Poland		Germany	
		Source	Span	Source	Span
Capital and growth	Investment growth rate	OECD.Stat	1991–2020	OECD.Stat	1984–2020
International economics	Trade-to-GDP ratio	OECD.Stat	1990–2020	OECD.Stat	1990–2020
Foreign trade	Net Export-to-GDP ratio	OECD.Stat	1990–2020	OECD.Stat	1984–2020
Monetary policy	Central bank discount rate	National Bank of Poland	1989–2020	Bundesbank	1984–2020
Business cycle	Output gap (HP filter)	OECD.Stat	1990–2020	OECD.Stat	1984–2020
Banking and credit	Credit-to-GDP ratio (actual)	BIS	1992–2020	BIS	1984–2020
Stock market (returns)	Growth rate of annual average stock market index	World Bank	1995–2020	World Bank	1989–2020
Stock market (volatility)	Average of the 360-day volatility of the national stock market index	World Bank	1995–2020	World Bank	1988–2020
Labour market	Unemployment rate	Statistics Poland	1990–2020	OECD.Stat	1984–2020
Energy sector	Primary energy production	Eurostat	1990–2020	Eurostat	1990–2020

Source: authors' own elaboration.

Table 4. Data for the model of oil shocks

Variable	Source	Span
Index of global real economic activity	Federal Reserve Bank of Dallas	1984–2020
Global crude oil production	U.S. Energy Information Administration (EIA)	1984–2020
U.S. crude oil imported acquisition cost by refiners (dollars per barrel)	U.S. Energy Information Administration (EIA)	1984–2020
U.S. CPI (all items) seasonally adjusted	OECD	1984–2020

Source: authors' own elaboration.

Table 5. Data transformations

Topic	Poland		Germany	
	Topic weight transformation	Economic indicator transformation	Topic weight transformation	Economic indicator transformation
Capital and growth	None	None	None	None
International economics	None	First difference	None	First difference
Foreign trade	None	First difference	None	First difference
Monetary policy	None	None	None	None
Business cycle	None	None	None	None
Crude oil market (oil supply shock)	None	None	None	None
Crude oil market (aggregate demand shock)	None	None	None	None
Crude oil market (oil specific-demand shock)	None	None	None	None
Banking and credit	None	First difference	None	First difference
Stock market (returns)	None	None	None	None
Stock market (volatility)	None	None	None	None
Labour market	None	First difference	None	First difference
Energy sector	None	First difference of logs	None	First difference of logs

Source: authors' own elaboration.

Table 6. Model details

Topic	Poland			Germany		
	VAR lag order	Deterministic terms	Standard errors	VAR lag order	Deterministic terms	Standard errors
Capital and growth	4	None	OLS	2	t	HAC
International economics	1	t, t^2	OLS	1	t, t^2	OLS
Foreign trade	3	t^2	HAC	4	t, t^2	HAC
Monetary policy	1	t, t^2	HAC	1	none	HAC
Business cycle	1	t, t^2	HAC	3	t, t^2	HAC
Crude oil market (oil supply shock)	1	t, t^2	HAC	1	none	OLS
Crude oil market (aggregate demand shock)	1	t, t^2	HAC	1	none	HAC
Crude oil market (oil specific-demand shock)	1	t, t^2	OLS	1	none	HAC
Banking and credit	1	none	OLS	1	none	OLS
Stock market (returns)	1	none	HAC	1	none	HAC
Stock market (volatility)	4	t, t^2	HAC	1	none	OLS
Labour market	1	none	OLS	2	t, t^2	OLS
Energy sector	1	none	HAC	4	t, t^2	HAC

Source: authors' own elaboration.