REGION

The Journal of ERSA
Powered by WU

Volume 11, Number 1, 2024, 55–78
DOI: 10.18335/region.v11i1.523

journal homepage: region.ersa.org
ISSN: 2409-5370

Bioeconomy firms and where to find them

Lukas Kriesch[1], **Sebastian Losacker**[1]

[1] Justus-Liebig-University Giessen, Giessen, Germany

Received: 15 November 2023/Accepted: 7 March 2024

Abstract. The bioeconomy represents a transformative approach to economic development and sustainability by harnessing biological resources and knowledge to produce goods, services, and energy while reducing dependence on non-renewable resources. In order to understand and support the bioeconomy, scholars and policymakers rely on an accurate measurement and monitoring of bio-based economic activities. However, existing statistical frameworks and industry classifications often fall short in capturing the unique characteristics of the bioeconomy. This article addresses this challenge by developing a methodological approach for comprehensive measurement and mapping of bio-based economic activities. We build a novel data set of bioeconomy firms in Germany using web-mining and machine learning techniques. This data set enables detailed analysis of bio-based economic activities, providing valuable insights into the spatial organization of the bioeconomy. The paper demonstrates the applicability of the data set by testing several hypotheses about the bioeconomy. Our research contributes to a better understanding of the bioeconomy's regional impacts and offers a valuable resource for policymakers and researchers interested in understanding the geography of bio-based economic activities. We make an aggregated version of the data set freely available online.

Key words: bioeconomy, web mining, natural language processing, regional data

1 Introduction

The bioeconomy represents a paradigm shift in our approach to economic development and sustainability. It encompasses the sustainable utilization of biological resources, such as plants, microorganisms, and biomass to produce a wide range of goods, services, and energy. The bioeconomy not only recognizes the potential of biological resources to meet our growing needs, but also acknowledges the need to reduce our dependence on non-renewable resources and to mitigate environmental impacts. By integrating novel sustainable technologies, innovative processes, and principles of circularity, the bioeconomy could offer a pathway towards a more sustainable and resilient future (Aguilar et al. 2018, Befort 2023, Bugge et al. 2016, Patermann, Aguilar 2021). Based on this vision, many countries have implemented a range of bioeconomy policies and strategies aiming to foster sustainable development (Prochaska, Schiller 2021, Proestou et al. 2023, Vogelpohl, Töller 2021).

The bioeconomy is also a promising concept for regional economies, as it offers regions the opportunity to diversify their economic base, foster innovation, and create new employment opportunities. By capitalizing on local biological resources and knowledge capabilities, regions can develop specialized clusters and value chains that leverage their unique ecological assets and knowledge bases (Kamath et al. 2023, Laasonen 2023, Martin et al. 2023, Morales, Dahlström 2022). In this context, the bioeconomy presents a viable pathway for regions to transition toward more sustainable economies. It can promote the adoption of sustainable practices, such as resource efficiency, waste valorization, and the circular economy, leading to reduced ecological footprints and enhanced regional sustainability.

However, it is important to note that the positive impacts of the bioeconomy, as well as its potential to foster regional development and sustainability, often remain speculative and are not guaranteed. The narrative of the bioeconomy as a universally beneficial approach is predominantly advocated by policymakers and related stakeholders who are keen on promoting its adoption. Yet, there is a growing body of scholarly work that raises critical questions regarding the assumed benefits of the bioeconomy (Allain et al. 2022, Bauer 2018, Bringezu et al. 2021, Friedrich et al. 2021). In that regard, economic geographers and regional scientists play a crucial role in analyzing the spatial dynamics of the bioeconomy, assessing its impacts on regional economies and informing policy interventions that foster sustainable regional development. Hence, scholars in economic geography and regional science can contribute to a better understanding of regional structural change and regional sustainability transitions towards a future bio-based economy.

Against this background, accurate tracking of bioeconomy activities is essential, not only for research purposes, but also for policymakers seeking to design effective strategies and place-based policies. Understanding the size, scope, and trends of bio-based economic activities provides policymakers with crucial insights into the bioeconomy's contribution to regional and national economies, job creation, and environmental sustainability. It enables them to identify emerging sectors, target support measures, and assess the effectiveness of policy interventions (El-Chichakli et al. 2016, Wesseler, von Braun 2017). However, measuring bio-based economic activities presents significant challenges. Traditional economic indicators often fail to capture the unique characteristics of the bioeconomy, such as the integration of biological resources and the circularity and sustainability of economic processes. Moreover, existing statistical frameworks, industry classifications and databases may lack comprehensive data on bioeconomy-related activities, making it difficult to obtain a complete and accurate picture. The multidimensional nature of the bioeconomy, spanning various sectors and encompassing both tangible and intangible elements, further complicates measurement efforts (Fischer et al. 2024, Losacker et al. 2023b, Ronzon et al. 2017, Wydra 2020).

In this paper, we contribute to solving these issues. The aim of this paper is to develop a methodological approach that allows a comprehensive measurement of bio-based economic activities. In that vein, we also aim to unveil the geography of bio-based economic activities. To this end, we build a unique dataset that enables us to identify and map bioeconomy firms in Germany. The dataset is based on a novel web-mining approach developed by Kriesch (2023). This dataset uses the open-source web repository CommonCrawl to identify German company websites and has proven to be a valuable database for spatial research. From this data, we identify bioeconomy firms using a combination of different natural language processing techniques, utilizing the semantic capabilities of modern transformer models (Reimers, Gurevych 2019, Vaswani et al. 2017). Our empirical approach allows for a detailed analysis of the economic activities of bio-based firms. That is to say, we are able to assess firms' technological capabilities and we can understand in which domains firms operate. In short, we establish a novel data source for monitoring the bioeconomy, overcoming several issues researchers and practitioners usually face when trying to measure bio-based economic activities. We test several hypotheses on the bioeconomy to validate our dataset and to demonstrate its applicability for future research, which is commonly done when introducing new methods or data to regional research (Abbasiharofteh et al. 2023, Ozgun, Broekel 2022). We make an aggregated version of our dataset freely accessible for fellow researchers, enabling further analyses and contributions to regional bioeconomy studies.

The remainder of this paper is organized as follows. In Section 2, we will introduce ideas behind the bioeconomy concept and review empirical findings that help to understand the geography of bio-based economic activities. This allows us to derive a couple of hypotheses about the bioeconomy and its geography. In Section 3, we explain our methodological approach in detail. We present our main results in Section 4. Section 5 concludes.

2 Literature review: What do we know about the bioeconomy?

2.1 The idea of a bio-based economy

The idea of a bioeconomy has emerged as a response to the need for sustainable development and the challenges posed by resource scarcity, climate change and environmental degradation. It refers to the (sustainable) utilization of biological resources to produce goods, services, and energy, and it encompasses diverse sectors such as agriculture, forestry, fisheries, biotechnology, and renewable energy. The use of bio-based products and technologies, however, is sought to span all economic sectors. The purpose of the bioeconomy is to shift from a linear economic model to a circular and regenerative approach, where biological resources are efficiently and responsibly managed (Aguilar et al. 2018, Allain et al. 2022, Befort 2023, Bugge et al. 2016, Patermann, Aguilar 2021).

For a transition towards a bio-based economy, the role of innovation emerges as paramount. Innovation, defined as any novel economic activity, is crucial to transition from a fossil economy to a bioeconomy (Befort 2023). As such, the bioeconomy encompasses a spectrum of innovation types, from drop-in solutions and bio-based substitutes to more transformative bio-based innovations that reshape socio-technical systems and redefine production networks (Befort 2023, Giurca, Befort 2023, Kuckertz et al. 2020, Losacker et al. 2023b). Nevertheless, the extent to which the bioeconomy and its innovations genuinely contribute to a more sustainable future remains somewhat ambiguous (Allain et al. 2022, Bauer 2018, Bringezu et al. 2021, Friedrich et al. 2021).

In the evolving discourse on the bioeconomy, there exists a multifaceted understanding of its implications and potential trajectories. In this regard, Bugge et al. (2016) delineate three distinct visions that encapsulate the breadth of scholarly perspectives on the bioeconomy. Firstly, the *biotechnology vision* underscores the pivotal role of biotechnology research, emphasizing its application and commercialization across diverse economic sectors. This vision is rooted in the belief that technological advancements and scientific progress can drive economic growth and innovation. Secondly, the *bio-resource vision* is anchored in the processing and enhancement of biological raw materials. It envisions a future where new value chains are established, leveraging the inherent potential of biological resources. Lastly, the *bio-ecology vision* emerges as a sustainability-centric perspective. It accentuates the importance of ecological processes that optimize energy and nutrient utilization, champion biodiversity, and caution against the pitfalls of monocultures and soil degradation. While expectations about actual sustainability outcomes of a future bioeconomy may vary according to these visions, there exists a broad consensus in the scholarly debate that the bioeconomy encompasses a wide array of industries and sectors (see above). This includes both traditional, low-tech goods and services in sectors such as forestry and agriculture, as well as more complex, knowledge-intensive economic activities like R&D in biotechnologies. In other words, the economic activities that can be associated with the bioeconomy are very diverse and thus difficult to track – and so is their geography.

2.2 The geography of bio-based economic activities

The bioeconomy emerges as a potent avenue for regional development, potentially enabling regional diversification, fostering regional innovation, and generating local employment opportunities. By utilizing local biological resources and leveraging regional knowledge capabilities, regions can establish specialized clusters and value chains that capitalize on their unique ecological assets and knowledge bases. In regional research, the bioeconomy has therefore received pronounced attention in recent years. For example, several researchers have delved into regional structural change and sustainability transitions toward a bio-based regional economy (Halonen et al. 2022, Laasonen 2023, Sanz-Hernández et al. 2019). Related studies have investigated innovative bio-clusters and regional innovation systems centered on biotechnologies (Abbasiharofteh, Broekel 2020, Heimeriks, Boschma 2014, Kamath et al. 2023), innovation networks (Bauer et al. 2018), and regional bioeconomy policies and strategies (Andersson, Grundel 2021). Additionally, there is increasing research interest in regional path creation and the pivotal role of actors and agency in propelling regional bioeconomy transitions (Martin et al. 2023, Morales, Dahlström 2022, Steinböck, Trippl 2023). Drawing from previous research findings on the bioeconomy, and complemented by insights from regional economics and economic geography, we can deduce several hypotheses concerning the geography of bio-based economic activities. In the following, we will enumerate these hypotheses and elucidate the rationale behind their formulation.

The bioeconomy, with its strong reliance on biomass and bio-based resources, raises the question of where bioeconomy firms tend to concentrate. There are compelling reasons to believe that bioeconomy enterprises tend to gravitate towards rural regions, resulting in the formation of rural bioeconomy clusters. These clusters may emerge as 'agricultural agglomerations' or 'Marshallian bio-districts' (Hermans 2021). Firstly, rural areas offer abundant biomass resources, including forests and agricultural products, providing a competitive advantage to bioeconomy firms reliant on biomass feedstocks. Secondly, rural regions often have established synergies with traditional industries like agriculture and forestry, offering infrastructure, knowledge, and expertise that bioeconomy firms can leverage for collaboration and innovation. Lastly, policy and regional development initiatives play a significant role in attracting bioeconomy firms to rural areas through financial incentives, grants, and supportive frameworks. These policies are often aimed at promoting rural development and can create a favorable investment climate for bioeconomy activities (Prochaska, Schiller 2024, Haarich, Kirchmayr-Novak Haarich, Kirchmayr-Novak). Based on these arguments, we assume that bioeconomy firms generally concentrate in rural areas, a hypothesis that is also supported by related empirical studies (Lasarte Lopez et al. 2023, Refsgaard et al. 2021).

Hypothesis 1 : *Given that the bioeconomy strongly relies on biomass and bio-based resources, bioeconomy firms concentrate in rural areas.*

In contrast to this first proposition, we argue that complex bioeconomy activities, such as those in biotechnology, concentrate in urban regions. Several arguments from the geography of innovation literature support this claim, building on core reasonings about agglomeration economies (Asheim et al. 2016, Broekel et al. 2023, Losacker et al. 2023a). Firstly, urban areas often provide a conducive environment for innovation and knowledge exchange. The concentration of universities, research institutions, and diverse talent pools in urban regions fosters collaboration, networking, and exchange of ideas. The availability of human capital and a strong regional innovation system attracts and supports the development of innovative bioeconomy activities. Secondly, urban regions typically have better access to specialized infrastructure and resources that are essential for cutting-edge research and development. Research facilities, laboratories, and technology parks are more prevalent in urban areas, providing infrastructure and equipment necessary for complex bioeconomy activities. Moreover, urban areas offer advanced transportation networks, communication systems, and logistical support, facilitating the movement of goods, services, and knowledge-intensive activities required for many innovative bioeconomy firms. Thirdly, urban regions often provide larger market opportunities and a diverse customer base. The concentration of various industries, markets, and consumers in urban areas creates a significant potential market for innovative bioeconomy products and services (Cooke 2002, Hermans 2018). Existing research supports this view, indicating that urban areas often host a higher concentration of complex bioeconomy activities (Ehrenfeld, Kropfhäußer 2017). In summary, while the hypothesis on bioeconomy firms concentrating in rural areas due to the strong reliance on biomass remains valid, there are additional reasons to claim that complex innovative bioeconomy activities, e.g., in biotechnology, concentrate in urban regions.

Hypothesis 2 : *Bioeconomy innovations and high-tech activities concentrate in urban areas.*

Next, we argue that economic activities centered on bio-based processes and biomass, such as activities in forestry or agriculture, typically locate in close proximity to their primary biomass feedstocks, a locational feature typical for bioeconomy clusters (Hermans 2021). This is underpinned by several reasons aligned with the basic principles of Weberian location theory. Firstly, being near the source of raw materials minimizes transportation costs, ensuring that the feedstock remains cost-effective for production or processing. Secondly, proximity to biomass sources ensures a consistent and timely supply, reducing potential downtimes or disruptions in the production process. Furthermore, being close to the source often means fresher inputs, which can be crucial for certain bio-based processes that rely on the quality and freshness of biomass. Lastly, such co-location fosters synergies with local agricultural or forestry sectors, promoting integrated value chains and facilitating efficient resource utilization. This geographical alignment between bio-based activities and their feedstock sources is not only economically prudent but also aligns with principles of sustainable production and consumption. The co-location of bioeconomy firms to biomass feedstocks is evident from several regional case studies on the bioeconomy (Martin et al. 2023, Martin, Coenen 2014, Ramirez 2021).

Hypothesis 3 : *Economic activities centered on bio-based processes and biomass locate in close proximity to their primary biomass feedstocks.*

We acknowledge that these hypotheses are somewhat generic. Nevertheless, their primary function is to highlight and showcase the novel dataset we have constructed in this paper, constituting the core contribution of our work.

3 Methods: Using web text data to map the bioeconomy

3.1 Issues in measuring the bioeconomy

We argue that previous attempts of measuring the bioeconomy are insufficient, mainly because the bioeconomy spans multiple sectors and can therefore not be captured using traditional methods or indicators. One of the challenges in measuring bio-based economic activities lies in the inadequacy of traditional statistical classifications, such as NACE and SIC codes or other sectoral classifications of economic activities, to fully capture the diverse nature of the bioeconomy. These classifications were primarily designed to categorize economic activities based on conventional industry sectors, often overlooking the unique characteristics and interconnections of bio-based economic activities. The bioeconomy, by its very nature, cuts across multiple sectors and involves a wide range of activities that may not neatly fit into traditional sectoral boundaries. For example, the bioeconomy includes sectors like biotechnology that span across different industries, combining elements of agriculture, manufacturing, and healthcare. It also encompasses activities like bio-energy production, bio-refineries, and bio-materials development, which do not align with conventional industry classifications (Jander, Grundmann 2019, Ronzon et al. 2017, Wesseler, von Braun 2017). Furthermore, the bioeconomy is characterized by innovation, continuous technological advancements, and the emergence of new value chains. Statistical classifications tend to be static and may struggle to keep pace with the dynamic and rapidly evolving nature of the bioeconomy. This dynamic nature often results in novel business models, cross-sector collaborations, and disruptive innovations that may not be adequately captured by existing statistical frameworks. These limitations are not only valid for measuring economic activities, but also for measuring innovation activities and knowledge generation where traditional indicators (e.g., patent data) also rely on statistical classifications that are not able to fully capture all parts of the bioeconomy (Fischer et al. 2024, Losacker et al. 2023b, Wydra 2020).

Prior efforts to gauge the bioeconomy have often relied on 'sector shares,' wherein researchers devise methodologies to determine the proportion of bio-based activities within traditional sector classifications like NACE (Lasarte Lopez et al. 2023, Ronzon et al. 2017). While this approach may be reasonable when assessing the bioeconomy at a national level, it is susceptible to what statisticians term an ecological fallacy when examined from a geographical standpoint. This fallacy involves making inferences about the characteristics of individual units (such as firms or regions) based on generalizations about the entire group (such as a nation). In many instances, researchers assume that, for example, 20% of the economic activity (value added, employment, etc.) in a specific sector is part of the bioeconomy. While this assumption may hold true on an aggregate level, it can lead to erroneous conclusions when applied to the firm or regional level. The share of bio-based activities within the sector may vary significantly across regions, being either notably higher or lower.

To overcome these limitations, researchers and policymakers need to explore alternative approaches that go beyond traditional statistical classifications. These approaches include adopting more flexible and adaptive frameworks that can capture the multi-dimensional and cross-sectoral aspects of the bioeconomy. Such frameworks may involve the development of new classification systems, the use of hybrid models that combine qualitative and quantitative data, and the integration of emerging indicators that reflect the unique characteristics of bio-based economic activities more adequately. In the next two sections (3.2 and 3.3), we propose an alternative way of measuring the bioeconomy. That is to say, we use a web-mining approach to retrieve information on firms from their website texts. Based on this text data, we employ machine learning techniques to identify, classify, and map bio-based economic activities.

3.2 A web-mining approach to identify bio-based economic activities in Germany

As data source we use website texts from German companies identified by Kriesch (2023). Our research focuses on the identification of bio-based economic activities in Germany, as Germany has emerged as a leading proponent of bioeconomy policies, recognizing its potential to drive sustainable economic growth and address environmental challenges (Imbert et al. 2017, Prochaska, Schiller 2021). At the regional level, Germany has implemented a range of policies and initiatives to promote the bioeconomy, supporting regional collaborations between businesses, research institutions, and policymakers to advance bio-based innovations. These policies aim to create favorable conditions for regional businesses to invest in bioeconomy activities, develop sustainable value chains, and contribute to regional economic development.

Web mining has witnessed significant advancements in recent years, primarily driven by the rise of natural language processing (NLP) techniques and the increasing digitization of data. Accordingly, the integration of web text data has become a vital complement to traditional data sources. In particular, unconventional information that may not be captured by traditional sources can be unveiled by means of web mining. Among other topics, companies use their websites to display information about products and services, their orientation and beliefs, strategies and relations with other companies (Gök et al. 2015). The dataset used in this paper, developed by Kriesch (2023), consists of 678,381 companies and their corresponding website texts. We updated the original dataset again in July 2023 to ensure that we have up-to-date information. We utilized geocoding techniques to map websites to geographical locations based on the information provided in their imprints. The address information was extracted using a fine-tuned named entity recognition model. We then used address geocoding to convert the address information into coordinates. In cases where a single company operates multiple domains, we attributed the content from each domain to the geographical location specified in its imprint. We extracted the HTML code of the first 25 subpages located on the landing page for each company domain, disregarding URLs predominantly comprised of machine-generated content and general legal information (e.g., imprint, cookie-policy, terms and conditions). Following Kinne, Lenz (2021), we argue that text located closer to the front or beginning of a domain is more likely to present general information about the whole company, while text positioned further back or towards the end tends to contain more specific details. Following the scraping process, the dataset comprises 9,601,260 subpages.

Text pre-processing plays a crucial role in converting raw HTML code into meaningful text data. Particularly, web data often contains a substantial amount of low quality and machine-generated content that may not be suitable for accurately predicting a company's capabilities. Hence, it is essential to employ appropriate data filtering and

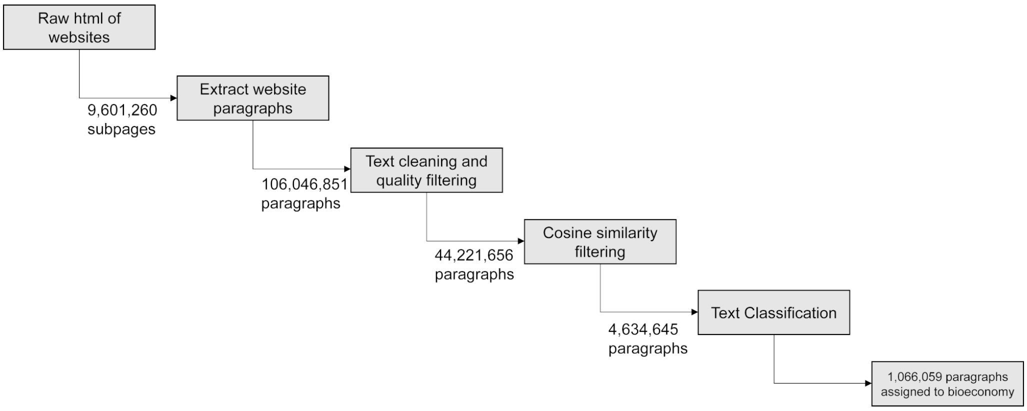
Raw html of websites
9,601,260 subpages
Extract website paragraphs
106,046,851 paragraphs
Text cleaning and quality filtering
44,221,656 paragraphs
Cosine similarity filtering
4,634,645 paragraphs
Text Classification
1,066,059 paragraphs assigned to bioeconomy

Figure 1: Flowchart of website text data processing

pre-processing techniques to mitigate the impact of irrelevant or misleading content on the predictive performance of the model. We illustrate the pre-processing procedure in Figure 1. In steps one and two, we extracted the main text from the HTML code, segmented the text into paragraphs and removed unwanted content like menus, headers, footers or advertisements. In the third filtering step, we applied further quality filters at the paragraph level to extract coherent and contextually well-embedded text. For this purpose, we use slightly adapted quality filtering heuristics developed by Rae et al. (2021). These heuristics have proven to be effective for the training of large language models and can be applied to prepare our dataset for analysis. Detailed information on the modifications made to these heuristics can be found in Appendix A.

We conducted a semantic search to identify paragraphs related to the bioeconomy. Therefore, we use a Sentence-BERT (SBERT) model to transform each of the extracted paragraphs into a complex numerical vector. SBERT models have demonstrated remarkable efficiency in leveraging the semantic knowledge of pre-trained transformer models, particularly when applied to downstream tasks such as semantic search or clustering. Their ability to capture and utilize semantic meanings has led to notable advancements in these specific applications, improving both accuracy and efficiency (Reimers, Gurevych 2019). We also embedded different keywords referring to the bioeconomy, as detailed in Appendix B, using the same SBERT model (German_Semantic_STS_V2). The choice of keywords for the semantic search and the methodology for manual annotation were thoroughly deliberated during a workshop involving bioeconomy experts. This collaborative session proved instrumental in gaining a nuanced understanding of what truly pertains to the bioeconomy. Unlike keyword-based searches that rely solely on matching specific terms, a semantic search employs advanced language understanding techniques to identify related and semantically similar concepts. This expanded scope of a semantic search enabled us to discover relevant content that may have been missed by a traditional keyword search. It facilitates the exploration of related ideas, synonyms, and contextually relevant information, leading to a more comprehensive and accurate retrieval of desired results. After calculating the cosine similarity between the vector representations of the paragraphs and keywords, we extracted those paragraphs that are related to the bioeconomy.

Figure 2 presents a density plot illustrating the distribution of cosine similarity values between the paragraphs and different exemplary keywords. To isolate scores indicating significant relevance, we computed the z-score corresponding to a two-tailed significance level of 0.01, yielding a critical value of approximately 2.576. This critical value helped delineate an upper threshold, represented by the red dashed lines on the histograms. Scores surpassing this threshold are considered statistically significant at the 0.01 level. Such high scores indicate texts with pronounced relevance to the selected keywords, marking them as candidates for in-depth examination. Following this filtering step, the dataset retained approximately 4.6 million paragraphs, constituting around 12.5 % of

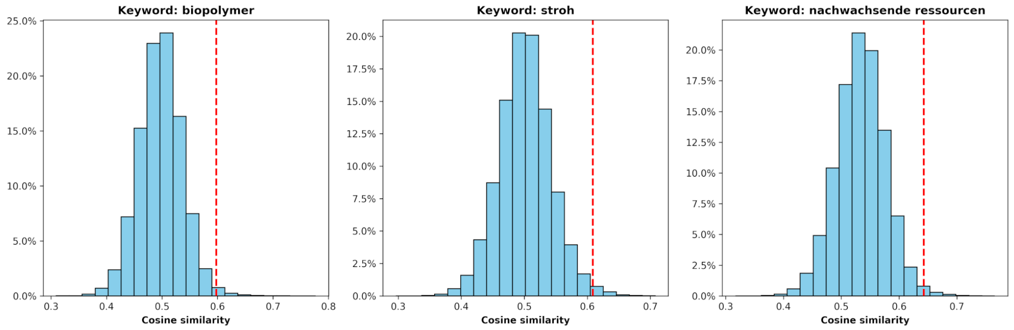

Figure 2: Distribution of cosine similarity values for selected keywords

the initial corpus.

In order to differentiate the technological capabilities of the extracted firms, we employ an advanced text classification approach. This approach builds upon the results obtained from semantic search and incorporates sophisticated techniques to accurately assess and distinguish the varying levels of technological capabilities exhibited by each firm. By leveraging this comprehensive text classification methodology, we can provide a more detailed and nuanced understanding of the technological landscape among the analyzed firms. To accomplish this, we extracted 1460 random paragraphs from the results of the semantic search. These paragraphs were then manually labeled to differentiate between three distinct levels of technological capabilities: (1) *No bioeconomy*, representing general information unrelated to the bioeconomy, (2) *bioeconomy in general*, encompassing sectors such as forestry and wood processing companies, (3) *bioeconomy high-tech*, comprising advanced and knowledge-intensive fields such as biotechnology and bio-pharmaceuticals. We provide a table with anchor examples from our annotation in Appendix C. This fine-grained labeling process allows for a detailed classification of the paragraphs, enabling us to assess effectively the technological capabilities of each firm within the bioeconomy domain.

For the training of the machine learning model, we again utilize a pre-trained SBERT model. Thanks to the inherent language understanding capabilities of those language models, we can fine-tune them with a relatively small number of manually annotated texts. By utilizing this transfer learning approach, we can efficiently adapt the models to our specific task of classifying technological capabilities in the bioeconomy domain, while minimizing the need for a large corpus of annotated data (Ruder et al. 2019). Given that a company's website typically consists of multiple subpages and paragraphs, it is plausible that different technological capabilities are assigned to a firm. To address this, we propose a straightforward heuristic where each company is labeled with the highest technological capability assigned to any of its relevant paragraphs. Consequently, we aim to provide a simplified but practical method for assigning labels to companies based on their most advanced technological capability as identified within their website content.

The dataset was divided into training, validation, and test sets to assess the performance of the model. After training the model on the training set and fine-tuning it using the validation set, we evaluated its accuracy on the test set. The achieved overall accuracy of 87.67 % indicates that the model was able to correctly classify the technological capabilities of the firms with a very high level of accuracy. The model has a precision of 88.68 % and a recall of 86.6 %, indicating its ability to accurately predict positive outcomes and capture true positives. The F1 score of 87.63 % demonstrates a well-balanced integration of precision and recall (Manning et al. 2008, Bishop 2006). By utilizing the knowledge and patterns learned during the training process, the model generated predictions for each paragraph, assigning them to one of the pre-defined classes representing different levels of technological capabilities within the bioeconomy domain.

Table 1: Description of the data set

Number of firms with a website	678,381
Number of bioeconomy firms (all)	142,949
Share of bioeconomy firms	21.07 %
Number of high-tech bioeconomy firms	13,554
Share of high-tech bioeconomy firms among all bioeconomy firms	9.48 %
Largest bioeconomy topics	Timber construction, agriculture, textiles
Average number of bioeconomy activities (topics) per firm	5.43

3.3 Uncovering different economic activities within the bioeconomy

While our previous efforts have focused on classifying bioeconomy firms based on their technological capabilities, there is an opportunity to delve deeper into the economic activities within the bioeconomy. This avenue of exploration can be effectively realized through the application of topic modeling techniques, which unveil underlying economic themes and patterns embedded within the dataset (Dahlke et al. 2024). To cluster the vector representations of the identified paragraphs, we utilize the BERTopic framework (Grootendorst 2022). Given the complexity of the vector embeddings, we implement a dimensionality reduction step. This process simplifies the vectors by distilling their information down to the most fundamental features. Such a simplification leads to more effective clustering, as it enhances the distinctiveness of the documents. We use UMAP for dimensionality reduction, which balances the preservation of essential local and global data structures (McInnes et al. 2018). Local data structures refer to the subtle relationships and patterns between neighboring data points, which are crucial for capturing the similarities between paragraphs. Global data structures, on the other hand, represent the broader distribution and relationships across the dataset, which are essential for preserving overarching thematic connections in the semantic space (McInnes et al. 2018). After dimensionality reduction, HDBSCAN is used for density-based clustering (McInnes et al. 2017). This technique effectively adapts to clusters of different shapes and densities, while effectively distinguishing between core topics and outliers. Finally, a modified version of TF-IDF, which focuses on clusters rather than documents, allows the identification of distinctive words that characterize each cluster.

The topic modeling methodology was extended to encompass all 1,066,059 paragraphs attributed to bioeconomy companies. In this pursuit, we chose a minimum cluster size of 1500, guided by the objective to construct clusters of meaningful coherence and comprehensive representation. In a concluding manual phase, we further clustered similar topics and subsequently refined the topic descriptions. In total, we extracted 55 topics from the corpus. For comprehensive insights into the results of our topic modeling analysis, please refer to Appendix E.

4 Results: Understanding the geography of bioeconomy firms

In total, we have identified 142,949 companies operating within the bioeconomy domain, of which 13,554 are classified as high-tech bioeconomy firms. Table 1 provides an overview of the dataset. Figure 3 shows the density distribution of bioeconomy firms in hexagonal cells.

4.1 The urban-rural divide in the bioeconomy

In order to gain insights into the geographical distribution of bioeconomy-related firms, we calculated the proportion of companies classified as either “bioeconomy” or “bioeconomy high-tech” within the entire cohort of observed web companies at the NUTS-3 level. Figure 4 shows the resulting map of firms active in the bioeconomy domain. We observe pronounced spatial disparities in the distribution of bioeconomy firms. In our overall findings, we observe a higher proportion of bioeconomy firms in regions characterized by lower population density. Regions with a high concentration of bioeconomy firms

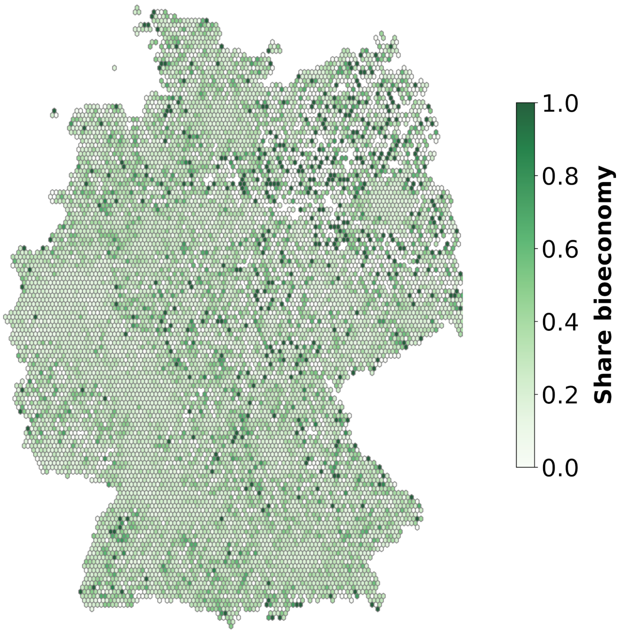

Figure 3: Share of bioeconomy firms

are, e.g., Südliche Weinstraße, Lüchow-Dannenberg, Oberallgäu, Regen or Cloppenburg (district).

Figure 4 also shows the bivariate relationship between the share of bioeconomy firms and population density. The inverse relationship observed between population density and bio-based firms aligns with the visual pattern depicted in the map. This negative correlation substantiates the initial visual impression, indicating a higher prevalence of bioeconomy firms in regions with lower population density.

To enhance the granularity of our findings, we classified the regions into distinct categories based on settlement types in Germany as defined by the BBSR (Federal Institute for Research on Building, Urban Affairs, and Spatial Development). Our results, as presented in Figure 4, reveal a discernible pattern that provides additional support for our earlier findings. Notably, urban centers have a smaller proportion of bioeconomy companies, while more thinly populated counties on average have a higher proportion of bioeconomy companies. To test the observed differences statistically, we ran a Welch-ANOVA, which confirmed that these differences are statistically significant ($p < 0.001$), with the exception of the comparison between “rural district with urbanization tendencies” and “sparsely populated rural district”. We employed a standard urban scaling framework (Bettencourt 2013, Broekel et al. 2023) as an alternative approach to scrutinize the geography of bio-based activities. Our findings reveal a scaling coefficient below one, signifying that bioeconomy firms are inclined to be situated in regions with a lower overall number of companies. Specifically, a 10 % surge in the number of companies within a given region corresponds to a 9.5 % increase in the number of bioeconomy firms, indicative of a sublinear scaling. This suggests that bioeconomy firms generally tend to thrive in less metropolitan areas. In summary, our results support the first hypothesis that bioeconomy firms are more likely to be found in rural regions.

4.2 The geography of high-tech bioeconomy firms

So far, we have shown the geography of bioeconomy companies in general. The following section discusses the geography of high-tech activities within the bioeconomy. Figure 5 shows the share of high-tech bioeconomic activities among all identified bioeconomic activities. Contrary to the general findings on the bioeconomy, high-tech activities are primarily concentrated in large cities. Districts with particularly large shares of high-tech activities are Jena, Heidelberg, Darmstadt and Munich. Moreover, Figure 5 provides an

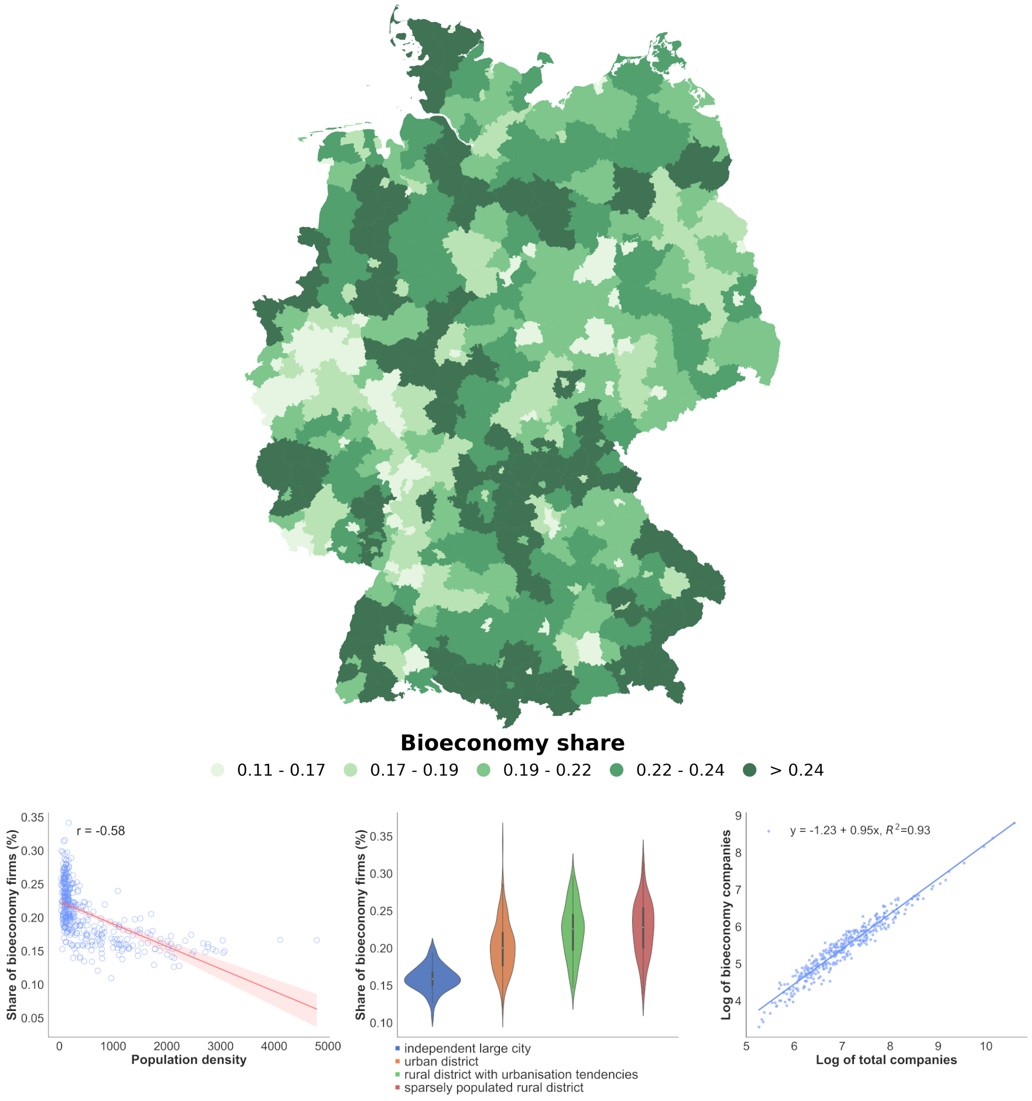

Figure 4: Geographical distribution of bioeconomy companies

indication of a positive linear correlation between the population density of a region and the share of high-tech bioeconomy firms. The variation in the share of high-tech bioeconomy firms between rural and urban regions is also well illustrated in Figure 5, which visualizes the distribution along settlement types. We find statistically significant ($p < 0.001$) differences between all pairwise comparisons except the comparison between “rural district with urbanization tendencies” and “sparsely populated rural district”. The scaling analysis reinforces these findings, revealing a superlinear scaling coefficient of 1.24. This signifies that a 10 % augmentation in the number of firms within a region corresponds to a 12.4 % increase in the number of high-tech bioeconomy firms in that region. The scaling coefficients of high-tech bioeconomy firms are remarkably similar to those derived from analyses using patent data (Bettencourt 2013, Broekel et al. 2023). This further strengthens the robustness of our findings, suggesting a consistent pattern of growth dynamics across different indicators of technological innovation and economic development. This observation suggests that high-tech bioeconomy firms thrive in metropolitan regions, propelled by urbanization economies. In conclusion, our results seem to indicate that the second hypothesis, namely the spatial concentration of high-tech firms in urban regions, is also empirically supported, although the results are less clear than in the analysis of the first hypothesis. This is due to a number of rural regions specializing in high-tech activities, as can be seen in Figure 5.

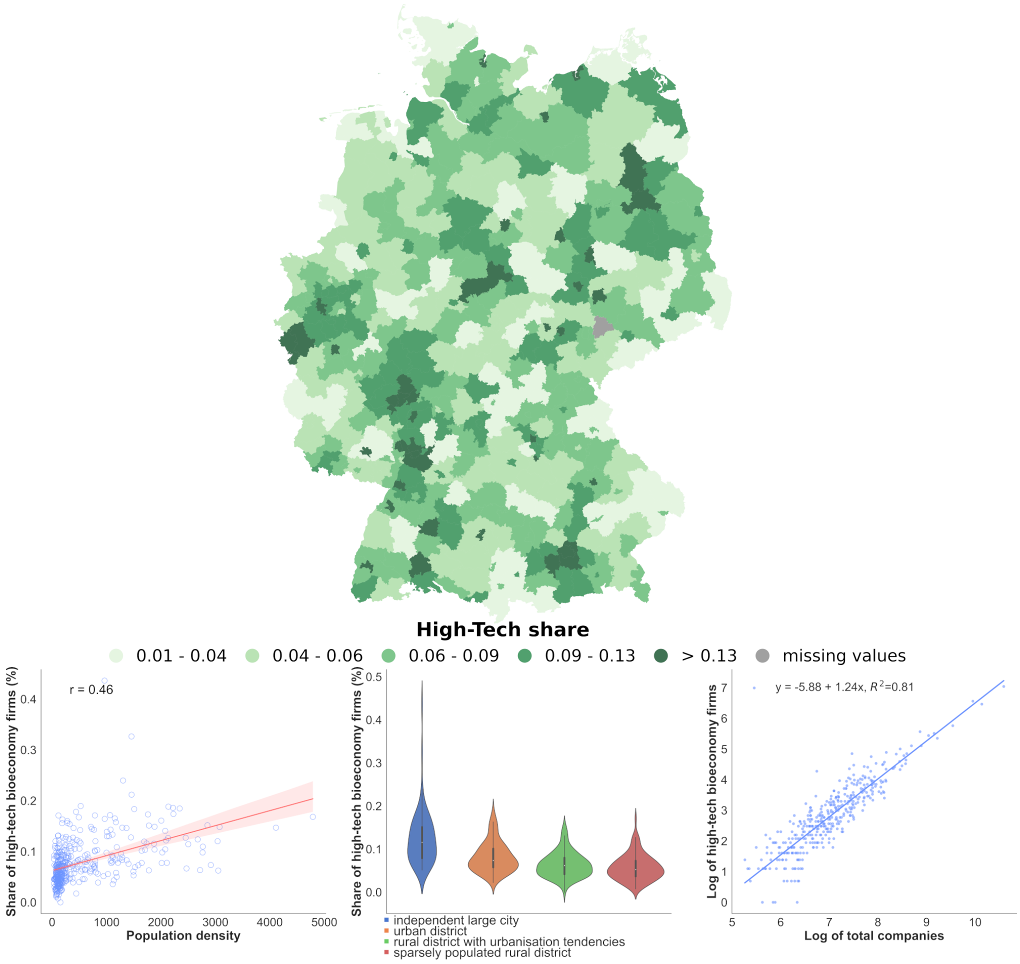

Figure 5: Geographical distribution of high-tech bioeconomy companies

4.3 The relationship between land use and economic activities in the bioeconomy

In the next step, we use the results of the topic modeling to gain further insights into the geography of bioeconomy firms in Germany. The topics identified provide information on the business segments in which a bioeconomy firm is active. The topics therefore go a step further than our delineation of high-tech companies that we used for the analysis of the second hypothesis. The topics should not be regarded as being directly equivalent to industry classifications, technology classes or economic sectors, but they do provide an indication of the business segments in which a company is active. The topics also provide an indication of which biogenic resources a firm uses for its business activities. We use this information to assess the third hypothesis, namely the spatial proximity of bioeconomy firms to their biomass needs. Figure 6 presents the correlation of the shares of different topics of all activities of the bioeconomy firms in a region with the share of the region's area that can be associated as a natural resource to a given topic. For example, one scatterplot illustrates the significant positive correlation of the activities of local bioeconomy companies related to “Wood” and the forest area of a region. The higher the share of forest area in a region, the higher the share of firms whose business activities are related to wood.

In conclusion, our analysis reveals that numerous hypotheses, which regional researchers might presuppose as applicable to the geography of bioeconomy firms, are indeed corroborated by the dataset we have compiled. We extended our analysis, as presented here, to the granularity of labor market regions, reinforcing the identified hypotheses. A corresponding set of figures for labor market regions is included in Appendix

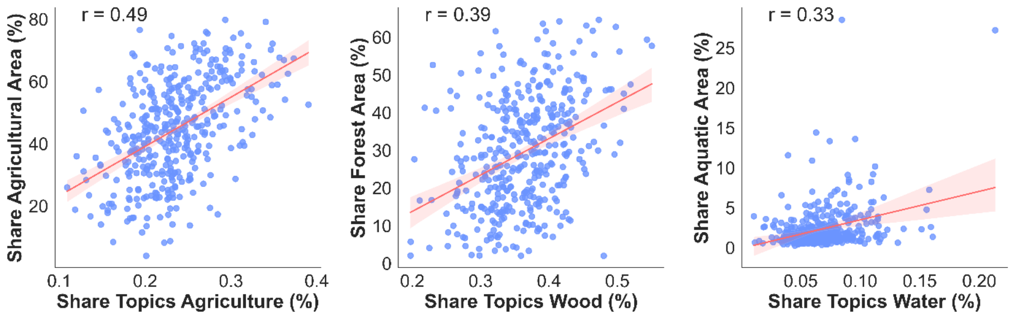

Figure 6: Land use and economic activities

D. As the main contribution of our work, we have presented a novel dataset, which will be a valuable resource for researchers focusing on the geography of bioeconomy activities. We have made an aggregated version of this dataset accessible in the online appendix. For each NUTS-3 region in Germany, the dataset encompasses several key variables, such as the total number of companies with a website, the count of firms identified as part of the bioeconomy, and metrics related to the various bioeconomy activities we have categorized. A comprehensive description of all variables is also provided in the online appendix. In utilizing the data generated in this paper, we urge researchers to contemplate the potential pitfalls associated with novel (web) data sources in regional research - a consideration familiar to many quantitative social scientists and economists (Einav, Levin 2014, Franklin 2022, 2023, Kitchin 2013).

The dataset can be utilized in diverse ways for quantitative research, such as linking bioeconomy activities with regional variables to understand the drivers of regional bioeconomy activities, following the increasing research on regional determinants of green economic activities (Losacker et al. 2023a). Alternatively, the data could serve as an independent variable to explore regional impacts, like the effect of local bioeconomy activities on regional development (e.g., value added, employment) or their correlation with environmental indicators (e.g., emissions, biodiversity loss). Qualitatively, the dataset can help identify regions for in-depth case studies, including those with notable bioeconomy activity or those with limited bioeconomy involvement.

5 Conclusion

This paper's main objective was to craft a methodological approach for the comprehensive measurement of bio-based economic activities, with an added focus on revealing their geographical distribution. This research goal was driven by the realization that traditional statistical classifications, be it industry or technology categories, fall short in encapsulating the nuances of bioeconomic activities. Such limitations not only deprive researchers of robust data for bioeconomy analysis, but also impede policymakers in their pursuit of evidence-informed decisions.

Against this background, we have built a unique dataset that enables us to identify and map bioeconomy firms in Germany. The dataset is based on a web-mining approach, using the open-source web repository CommonCrawl to identify German company websites. From this data, we have identified bioeconomy firms using a combination of different natural language processing techniques, utilizing the semantic capabilities of modern transformer models. Our final dataset enables a precise analysis of the bioeconomy, its geography, and its various domains. We used this dataset to test three hypotheses about the bioeconomy, thereby assessing the applicability of the dataset and demonstrating its potential for empirical research. First, we showed that bioeconomy firms predominantly concentrate in rural areas. Second, however, we demonstrated that high-tech activities related to the bioeconomy concentrate in urban areas. Third, we found that economic activities centered on bio-based processes and biomass locate in close proximity to their primary biomass feedstocks.

However, it is important to interpret these results with caution and keep in mind several limitations. The presence of a firm's website varies significantly based on specific firm attributes. Consequently, our suggested web mining framework may not be applicable for analyzing certain firms. In particular, firms that are either very new or very small, along with those operating in specific sectors and regions, tend to have limited website availability (see Kinne, Axenbeck 2020). For our study, this implies a particular gap in information regarding small agricultural firms located in rural areas, which are underrepresented due to these constraints. Furthermore, webpage data is subject to a self-description bias, since firms have the autonomy to choose both the nature and the manner in which information is presented on their sites.

We provide the compiled dataset in an aggregated form along with variable descriptions in the online appendix. We encourage fellow researchers to utilize this dataset to address the numerous unresolved research questions surrounding the bioeconomy. We are confident that future analyses of this data will yield important insights, paving the way for place-specific recommendations that can inform industrial and innovation policies geared towards sustainable regional development.

Acknowledgements

We acknowledge financial support by the German Federal Ministry of Education and Research (BMBF 031B1281). Previous versions of this article were presented at the 16th Rauischholzhausen symposium on Economic Geography (April 2023), at the RSA Annual Conference in Ljubljana (June 2023), at an annual meeting of junior bioeconomy research groups funded by the BMBF (November 2023), and at the 7th Geography of Innovation Conference in Manchester (January 2024). The authors have benefited greatly from feedback received from the participants of these events. The methodological approach employed in this paper has been discussed during a workshop on advanced methods for the measurement of economic and innovation activities in the bioeconomy, held at Giessen University in August 2023. We thank the discussants, in particular Henning Kroll, Daniel Schiller and Sven Wydra, for feedback during this workshop.

Declaration of interest

No potential competing interest was reported by the authors.

References

Abbasiharofteh M, Broekel T (2020) Still in the shadow of the wall? The case of the Berlin biotechnology cluster. *Environment and Planning A: Economy and Space*: 0308518X2093390. CrossRef

Abbasiharofteh M, Krüger M, Kinne J, Lenz D, Resch B (2023) The digital layer: Alternative data for regional and innovation studies. *Spatial Economic Analysis* 18: 507–529. CrossRef

Aguilar A, Wohlgemuth R, Twardowski T (2018) Perspectives on bioeconomy. *New Biotechnology* 40: 181–184. CrossRef

Allain S, Ruault JF, Moraine M, Madelrieux S (2022) The 'bioeconomics vs bioeconomy' debate: Beyond criticism, advancing research fronts. *Environmental Innovation and Societal Transitions* 42: 58–73. CrossRef

Andersson I, Grundel I (2021) Regional policy mobilities: Shaping and reshaping bioeconomy policies in Värmland and Västerbotten, Sweden. *Geoforum* 121: 142–151. CrossRef

Asheim B, Grillitsch M, Trippl M (2016) Regional innovation systems: Past – presence – future. In: Shearmur R, Carrincazeaux C, Doloreux D (eds), *Handbook on the geographies of innovation*. Edward Elgar Publishing Ltd., 45–62. CrossRef

Bauer F (2018) Narratives of biorefinery innovation for the bioeconomy: Conflict, consensus or confusion? *Environmental Innovation and Societal Transitions* 28: 96–107. CrossRef

Bauer F, Hansen T, Hellsmark H (2018) Innovation in the bioeconomy–dynamics of biorefinery innovation networks. *Technology Analysis and Strategic Management* 30: 935–947. CrossRef

Befort N (2023) *The Bioeconomy: Institutions, Innovation and Sustainability*. Routledge. CrossRef

Bettencourt LM (2013) The origins of scaling in cities. *Science* 340: 1438–1441. CrossRef

Bishop C (2006) *Pattern Recognition and Machine Learning*. Springer New York. CrossRef

Bringezu S, Distelkamp M, Lutz C, Wimmer F, Schaldach R, Hennenberg KJ, Böttcher H, Egenolf V (2021) Environmental and socioeconomic footprints of the German bioeconomy. *Nature Sustainability* 4: 775–783. CrossRef

Broekel T, Knuepling L, Mewes L (2023) Boosting, sorting and complexity – urban scaling of innovation around the world. *Journal of Economic Geography*. CrossRef

Bugge M, Hansen T, Klitkou A (2016) What is the bioeconomy? A review of the literature. *Sustainability* 8: 691. CrossRef

Cooke P (2002) Biotechnology clusters as regional, sectoral innovation systems. *International Regional Science Review* 25: 8–37. CrossRef

Dahlke J, Beck M, Kinne J, Lenz D, Dehghan R, Wörter M, Ebersberger B (2024) Epidemic effects in the diffusion of emerging digital technologies: Evidence from artificial intelligence adoption. *Research Policy* 53: 104917. CrossRef

Ehrenfeld W, Kropfhäußer F (2017) Plant-based bioeconomy in Central Germany – A mapping of actors, industries and places. *Technology Analysis & Strategic Management* 29: 514–527. CrossRef

Einav L, Levin J (2014) Economics in the age of big data. *Science* 346: 6210. CrossRef

El-Chichakli B, von Braun J, Lang C, Barben D, Philp J (2016) Policy: Five cornerstones of a global bioeconomy. *Nature* 535: 221–223. CrossRef

Fischer L, Losacker S, Wydra S (2024) National specialization and diversification in the bioeconomy: Insights from biobased technologies in chemical and pharmaceutical sectors. *Technology in Society* 76: 102462. CrossRef

Franklin R (2022) Quantitative methods I: Reckoning with uncertainty. *Progress in Human Geography* 46: 689–697. CrossRef

Franklin R (2023) Quantitative methods II: Big theory. *Progress in Human Geography* 47: 178–186. CrossRef

Friedrich J, Bunker I, Uthes S, Zscheischler J (2021) The potential of bioeconomic innovations to contribute to a social-ecological transformation: A case study in the livestock system. *Journal of Agricultural and Environmental Ethics* 34: 1–26. CrossRef

Giurca A, Befort N (2023) Deconstructing substitution narratives: The case of bioeconomy innovations from the forest-based sector. *Ecological Economics* 207: 107753. CrossRef

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. https://arxiv.org/abs/2203.05794v1

Gök A, Waterworth A, Shapira P (2015) Use of web mining in studying innovation. *Scientometrics* 102: 653–671. CrossRef

Haarich S, Kirchmayr-Novak S European Commission, Joint Research Centre, Publications Office of the European Union, Luxembourg. CrossRef

Halonen M, Näyhä A, Kuhmonen I (2022) Regional sustainability transition through forest-based bioeconomy? Development actors' perspectives on related policies, power, and justice. *Forest Policy and Economics* 142: 102775. CrossRef

Heimeriks G, Boschma R (2014) The path- and place-dependent nature of scientific knowledge production in biotech 1986-2008. *Journal of Economic Geography* 14: 339–364. CrossRef

Hermans F (2018) The potential contribution of transition theory to the analysis of bioclusters and their role in the transition to a bioeconomy. *Biofuels, Bioproducts and Biorefining* 12: 265–276. CrossRef

Hermans F (2021) Bioclusters and sustainable regional development. In: Sedita SR, Blasi S (eds), *Rethinking Clusters: Place-based Value Creation in Sustainability Transitions*. Springer, Cham, 81–91. CrossRef

Imbert E, Ladu L, Morone P, Quitzow R (2017) Comparing policy strategies for a transition to a bioeconomy in Europe: The case of Italy and Germany. *Energy Research & Social Science* 33: 70–81. CrossRef

Jander W, Grundmann P (2019) Monitoring the transition towards a bioeconomy: A general framework and a specific indicator. *Journal of Cleaner Production* 236. CrossRef

Kamath R, Elola A, Hermans F (2023) The green-restructuring of clusters: Investigating a biocluster's transition using a complex adaptive system model. *European Planning Studies* 31: 1842–1867. CrossRef

Kinne J, Axenbeck J (2020) Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics* 125: 2011–2041. CrossRef

Kinne J, Lenz D (2021) Predicting innovative firms using web mining and deep learning. *PLOS ONE* 16: e0249071. CrossRef

Kitchin R (2013) Big data and human geography. *Dialogues in Human Geography* 3: 262–267. CrossRef

Kriesch L (2023) *Web Mining und Natural Language Processing als methodisches Komplement in der Wirtschaftsgeographie*. CrossRef

Kuckertz A, Berger ES, Brändle L (2020) Entrepreneurship and the sustainable bioeconomy transformation. *Environmental Innovation and Societal Transitions* 37: 332–344. CrossRef

Laasonen V (2023) Building dynamic capabilities in the transition toward a knowledge-based bioeconomy: A case study of three Finnish regions. *Regional Studies*: 1–12. CrossRef

Lasarte Lopez J, González Hermoso H, Rossi Cervi W, Van Leeuwen M, M‘barek R (2023) BioRegEU. a pilot dataset for regional employment and value added in the EU bioeconomy. Publications office of the european union. CrossRef

Losacker S, Hansmeier H, Horbach J, Liefner I (2023a) The geography of environmental innovation: A critical review and agenda for future research. *Review of Regional Research*: 1–26. CrossRef

Losacker S, Heiden S, Liefner I, Lucas H (2023b) Rethinking bioeconomy innovation in sustainability transitions. *Technology in Society* 74: 102291. CrossRef

Manning CD, Raghavan P, Schütze H (2008) *Introduction to Information Retrieval*. Cambridge University Press. CrossRef

Martin H, Coenen L (2014) Institutional context and cluster emergence: The biogas industry in Southern Sweden. *European Planning Studies* 23: 2009–2027. CrossRef

Martin H, Grundel I, Dahlström M (2023) Reconsidering actor roles in regional innovation systems: Transformative industrial change in the forest-based bioeconomy. *Regional Studies* 57: 1636–1648. CrossRef

McInnes L, Healy J, Astels S (2017) hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software* 2: 205. CrossRef

McInnes L, Healy J, Melville J (2018) UMAP: Uniform manifold approximation and projection for dimension reduction. *The Journal of Open Source Software* 3: 861. CrossRef

Morales D, Dahlström M (2022) Smart specialization and participatory processes in green path renewal. Analysis of the forest-based bioeconomy in sparsely populated regions in the Nordics. *European Planning Studies*: 1–20. CrossRef

Ozgun B, Broekel T (2022) Assessing press releases as a data source for spatial research. *REGION* 9: 25–44. CrossRef

Patermann C, Aguilar A (2021) A bioeconomy for the next decade. *EFB Bioeconomy Journal* 1: 100005. CrossRef

Prochaska L, Schiller D (2021) An evolutionary perspective on the emergence and implementation of mission-oriented innovation policy: The example of the change of the leitmotif from biotechnology to bioeconomy. *Review of Evolutionary Political Economy* 2: 141–249. CrossRef

Prochaska L, Schiller D (2024) Spatial distribution of bioeconomy R&D funding: Opportunities for rural and lagging regions? *European Planning Studies*: 1–21. CrossRef

Proestou M, Schulz N, Feindt PH (2023) A global analysis of bioeconomy visions in governmental bioeconomy strategies. *Ambio 2023*: 1–13. CrossRef

Rae JW, Borgeaud S, Cai T, Millican K, Hoffmann J, Song F, Aslanides J, Henderson S, Ring R, Young S, Rutherford E, Hennigan T, Menick J, Cassirer A, Powell R, Driessche GVD, Hendricks LA, Rauh M, Huang PS, Glaese A, Welbl J, Dathathri S, Huang S, Uesato J, Mellor J, Higgins I, Creswell A, Mcaleese N, Wu A, Elsen E, Jayakumar S, Buchatskaya E, Budden D, Sutherland E, Simonyan K, Paganini M, Sifre L, Martens L, Li L, Kuncoro A, Nematzadeh A, Gribovskaya E, Donato D, Lazaridou A, Mensch A, Lespiau JB, Tsimpoukelli M, Grigorev N, Fritz D, Sottiaux T, Pajarskas M, Pohlen T, Gong Z, Toyama D, D'autume CDM, Li Y, Terzi T, Mikulik V, Babuschkin I, Clark A, De D, Casas L, Guy A, Jones C, Bradbury J, Johnson M, Hechtman B, Weidinger L, Gabriel I, Isaac W, Lockhart E, Osindero S, Rimell L, Dyer C, Vinyals O, Ayoub K, Stanway J, Bennett L, Hassabis D, Kavukcuoglu K, Irving G (2021) Scaling language models: Methods, analysis & insights from training Gopher. https://arxiv.org/abs/2112.11446v2

Ramirez P (2021) Technological revolutions, socio-technical transitions and the role of agency: Värmland's transition to a regional bio-economy. *Regional Studies* 55: 1642–1651. CrossRef

Refsgaard K, Kull M, Slätmo E, Meijer MW (2021) Bioeconomy – A driver for regional development in the Nordic countries. *New Biotechnology* 60: 130–137. CrossRef

Reimers N, Gurevych I (2019) Sentence-BERT: Sentence embeddings using Siamese BERT-networks. https://github.com/UKPLab/

Ronzon T, Piotrowski S, M'Barek R, Carus M (2017) A systematic approach to understanding and quantifying the EU's bioeconomy. *Bio-based and Applied Economics* 6: 1–17. CrossRef

Ruder S, Peters ME, Swayamdipta S, Wolf T (2019) Transfer learning in natural language processing. *Proceedings of the 2019 Conference of the North*: 15–18. CrossRef

Sanz-Hernández A, Sanagustín-Fons MV, López-Rodríguez ME (2019) A transition to an innovative and inclusive bioeconomy in Aragon, Spain. *Environmental Innovation and Societal Transitions* 33: 301–316. CrossRef

Steinböck N, Trippl M (2023) The thorny road towards green path development: The case of bioplastics in Lower Austria. *Regional Studies, Regional Science* 10: 735–749. CrossRef

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Łukasz Kaiser, Polosukhin I (2017) Attention is all you need. *Advances in Neural Information Processing Systems* 30. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Vogelpohl T, Töller AE (2021) Perspectives on the bioeconomy as an emerging policy field. *Journal of Environmental Policy & Planning* 23: 143–151. CrossRef

Wesseler J, von Braun J (2017) Measuring the bioeconomy: Economics and policies. *Annual Review of Resource Economics* 9: 275–298. CrossRef

Wydra S (2020) Measuring innovation in the bioeconomy — Conceptual discussion and empirical experiences. *Technology in Society* 61: 101242. CrossRef

A Appendix: Text quality filtering

We adapted the text quality filtering heuristics established by Rae et al. (2021) to better suit the nuances of the German language, implementing the following modifications to the filtering process:

- Paragraphs are excluded if the average word length falls outside the 3 to 12 character range.
- Paragraphs are eliminated if the ratio of symbols to words exceeds 0.15.
- Paragraphs are discarded if they contain fewer than two stopwords from either English or German.
- Any paragraph composed entirely of uppercase letters is also removed from consideration, as this often signifies non-standard text or spam.

B Appendix: Keyword list

Table B.1: Keyword list

biobasiert	Biokraftstoff	Naturfasern	Agrar	Bioökonomie
Holz	Biotenside	Biopharmazeutika	Aquakulturen	Landwirtschaft
Zellstoff	Bioschmierstoff	Mikroorganismen	Mikrobiom	Aquaponik
Biophysik	Bioengineering	Enzym	Biolösungsmittel	Peptide
Biotechnik	Biomasse	Biokunststoff	Biopolymer	Nukleoside
Biotechnologie	nachwachsende Rohstoffe	Bioplastik	biologisch abbaubar	Vertikale Landwirtschaft
Bioenergie	biogen	Pflanze	Papier	Nachwachsende Ressourcen
Biochemie	biologisch	Forst	Fischerei	Bioklebstoff
Biokosmetik	pflanzenbasiert			

C Appendix: Anchor examples of annotated training data

Table C.2: Anchor examples of annotated training data

Label	English (translated)	German (original)
Bioeconomy general	HOLZ-BARAN is your partner for woodworking, floor coverings and solid wood flooring in Luppa near Leipzig in Saxony.	HOLZ-BARAN ist Ihr Partner rund um Holzbearbeitung, Bodenbeläge und Massivholzdielen in Luppa bei Leipzig in Sachsen.
Bioeconomy general	Raiffeisen Bio-Brennstoffe GmbH sells wood chips, biomass, wood briquettes and, in particular, wood pellets. Together with co-operative partners, the associated company of AGRAVIS Raiffeisen AG has built up an efficient sales network throughout the northern half of Germany.	Die Raiffeisen Bio-Brennstoffe GmbH vertreibt Hackschnitzeln, Biomasse, Holzbriketts und insbesondere Holzpellets. Gemeinsam mit genossenschaftlichen Partnern hat das Beteiligungsunternehmen der AGRAVIS Raiffeisen AG ein leistungsfähiges Vertriebsnetz in der gesamten Nordhälfte Deutschlands aufgebaut.
Bioeconomy general	Our meat factory consists of several organic farms, all of which are located in our beautiful district of Höxter in East Westphalia. In the easternmost corner of NRW, where there are still many meadows and pastures!	Unsere Fleischmanufaktur besteht aus mehrere Bio-Höfen, die alle in unserem schönem Kreis Höxter in Ostwestfalen angesiedelt sind. Im östlichsten Fleck von NRW, wo es noch viele Wiesen & Weiden gibt!

Table C.2: Anchor examples of annotated training data - continued

Label	English (translated)	German (original)
Bioeconomy high-tech	We are BioCer Entwicklungs-GmbH from Bayreuth. As a young, innovative medical technology company, we specialise in the research, development and production of innovative medical products made from biomaterials without animal or human components. As a service provider, we also coat implants for our customers and develop new types of medical products.	Wir sind die BioCer Entwicklungs-GmbH aus Bayreuth. Als junges, innovatives Unternehmen der Medizintechnik haben wir uns auf die Forschung, Entwicklung und Produktion von innovativen Medizinprodukten aus Biomaterialien ohne tierische oder humane Bestandteile spezialisiert. Darüber hinaus beschichten wir als Dienstleister für unsere Kunden Implantate und entwickeln neuartige Medizinprodukte.
Bioeconomy high-tech	Hansen is a global biotechnology company that develops natural solutions for the food, nutrition, pharmaceutical and agricultural industries.	Hansen ist ein globales Biotechnologieunternehmen, das natürliche Lösungen für die Lebensmittel-, Ernährungs-, Pharma- und Landwirtschaftsindustrie entwickelt.
Bioeconomy high-tech	The BMW Group is developing innovative, bio-based surfaces in cooperation with start-up companies. For example, the newly developed DeserttexTM is made from powdered cactus fibres and a bio-based polyurethane matrix. This allows the elimination of animal-based raw materials to be combined with a reduction in CO2 emissions.	Die BMW Group entwickelt in Kooperation mit Start-up-Unternehmen innovative, biobasierte Oberflächen. So setzt sich z. B. das neuentwickelte DeserttexTM aus pulverisierten Kaktusfasern und einer biobasierten Polyurethan-Matrix zusammen. So lässt sich der Verzicht auf tierische Rohstoffe mit einer Co2-Reduzierung kombinieren.
No bioeconomy	Passport photos are subject to strict regulations (biometric photos). We are always familiar with the latest standards in order to be able to offer you successful passport photos at all times.	Passbilder unterliegen strengen Vorschriften (Biometrische Fotos). Wir sind immer mit den neuesten Standards vertraut, um Ihnen jederzeit gelungene Passbilder bieten zu können.
No bioeconomy	Today, slate is quarried underground and above ground using innovative and technically advanced processing methods in an environmentally friendly way with the aid of state-of-the-art technology. Efficient laying techniques make this sustainable natural product extremely economical.	Schiefer wird heute durch innovative und technisch weiterentwickelte Bearbeitungsmethoden umweltschonend mit Hilfe modernster Technik unter und über Tage abgebaut. Rationelle Verlegetechniken machen das nachhaltige Naturprodukt äußerst wirtschaftlich.

D Appendix: Findings for labor market regions

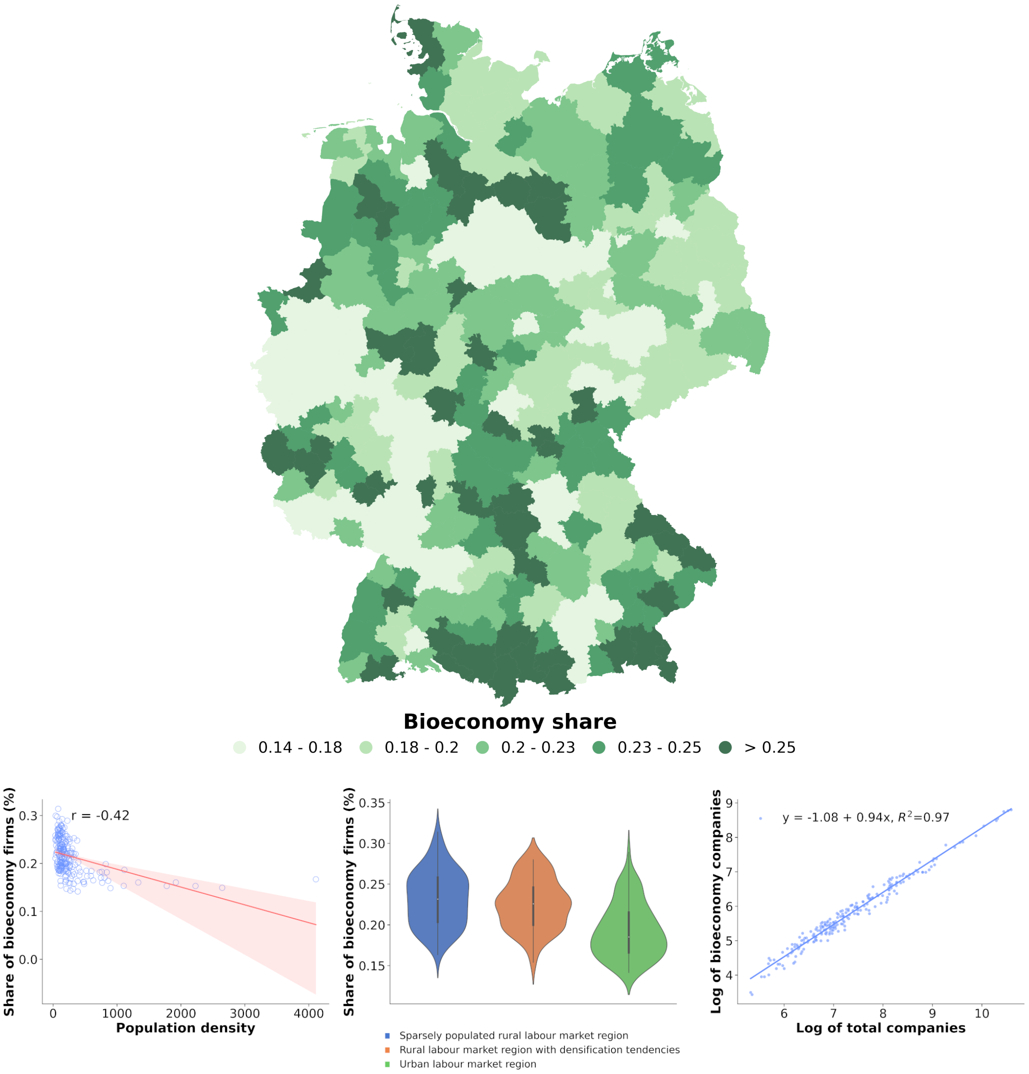

Figure D.1: Geographical distribution of bioeconomy companies (adapted version of Figure 4)

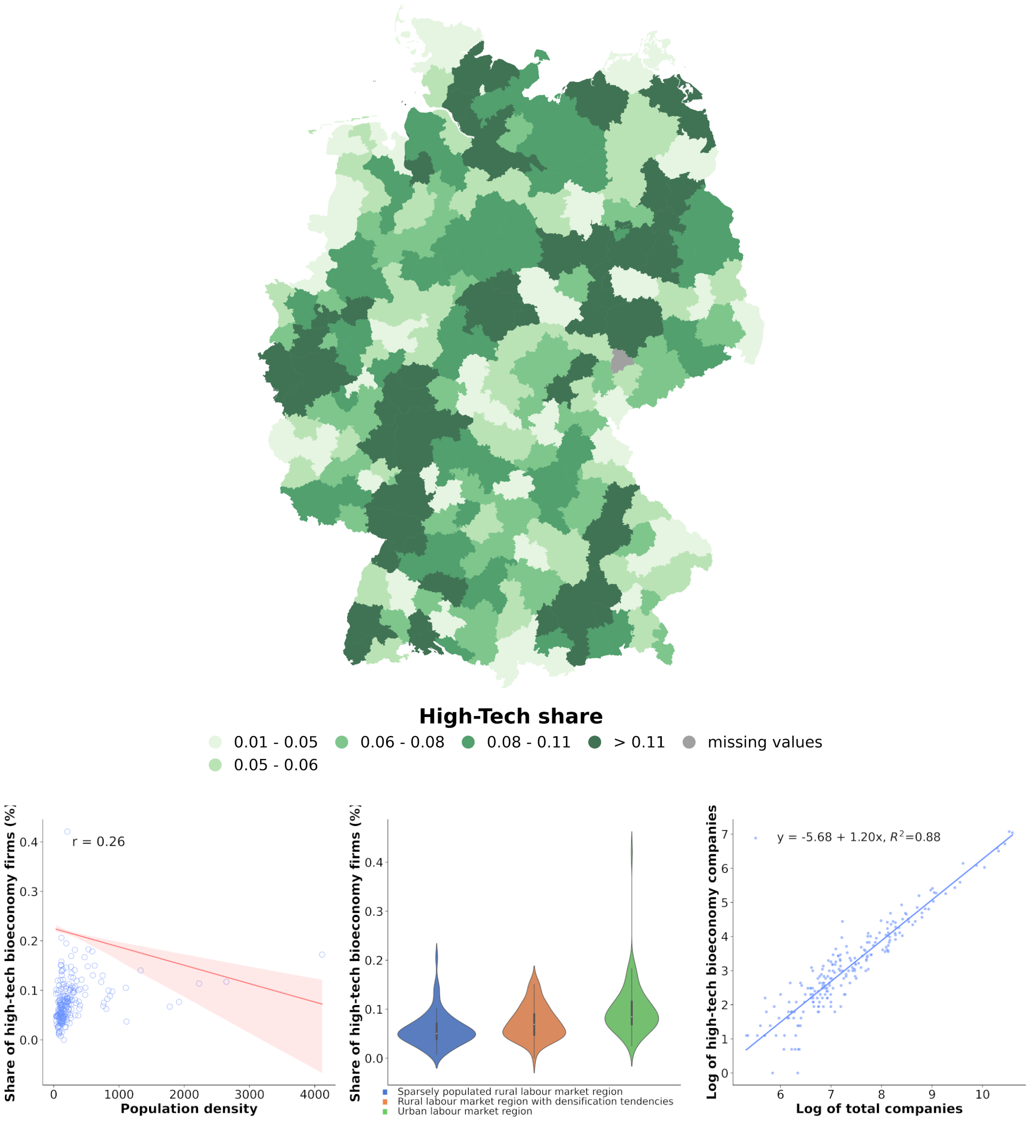
High-Tech share
● 0.01 - 0.05
● 0.05 - 0.06
● 0.06 - 0.08
● 0.08 - 0.11
● > 0.11
● missing values
Share of high-tech bioeconomy firms (%)
r = 0.26
0.0
0.1
0.2
0.3
0.4
0
1000
2000
3000
4000
Population density
Share of high-tech bioeconomy firms (%)
0.0
0.1
0.2
0.3
0.4
Sparsely populated rural labour market region
Rural labour market region with densification tendencies
Urban labour market region
Log of high-tech bioeconomy companies
y = -5.68 + 1.20x, R^2=0.88
0
1
2
3
4
5
6
7
6
7
8
9
10
Log of total companies

Figure D.2: Geographical distribution of bioeconomy companies (adapted version of Figure 5)

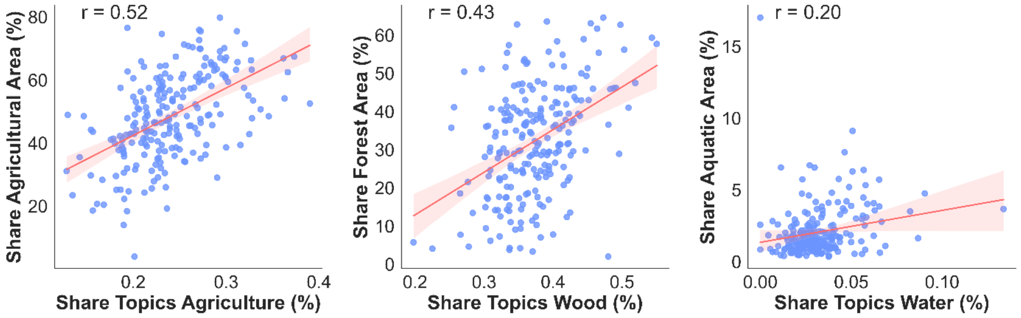

Figure D.3: Land use and economic activities (adapted version of Figure 6)

E Appendix: Topic model based on bioeconomy paragraphs

Table E.1: Topic model based on bioeconomy paragraphs

Topic	Name	Top words	Share
1	wood	holz', 'möbel', 'holzbau', 'bauen', 'materialien', 'haus', 'massivholz'	10.0 %
2	agriculture	bio', 'produkte', 'gemüse', 'landwirtschaft', 'obst', 'lebensmittel', 'qualität', 'region', 'ernährung'	8.5 %
3	miscellaneous	'products', 'research', 'food', 'high', 'well', 'quality', 'development', 'new', 'use', 'based'	7.6 %
4	textiles	baumwolle', 'leder', 'wolle', 'materialien', '100', 'material', 'farben', 'cm', 'teppiche'	6.3 %
5	agriculture	pflanzen', 'garten', 'blumen', 'blüten', 'rosen', 'stauden', 'pflege', 'pflanze', 'balkon'	5.2 %
6	wood	wald', 'bäume', 'wälder', 'natur', 'baum', 'mehr', 'co2', 'flächen'	4.9 %
7	cosmetics & food supplements	'haut', 'wirkung', 'inhaltsstoffe', 'haar', 'duft', 'aloe', 'vera', 'wirkt', 'naturkosmetik', 'öle'	3.8 %
8	food	fleisch', 'wurst', 'geschmack', 'steak', 'qualität', 'wurstwaren', 'rind', 'metzgerei', 'rindfleisch'	2.8 %
9	biotechnology	entwicklung', 'forschung', 'unternehmen', 'bereich', 'biotechnologie', 'dr', 'universität', 'entwickelt'	2.6 %
10	wood	holz', 'cm', 'kinder', 'gefertigt', 'material','spielzeug'	2.5 %
11	pulp and paper	papier', 'fsc', 'wellpappe', 'verpackungen', 'verpackung', 'rohstoffen', '100', 'biologisch', 'nachwachsenden'	2.3 %
12	animal feed	'futter', 'hunde', 'pferde', 'hund', 'pferd', 'fütterung', 'ernährung', 'vitamine'	2.2 %
13	wood	'pellets', 'holzpellets', 'heizen', 'holz', 'brennholz', 'brennstoff'	2.0 %
14	wood	parkett', 'laminat', 'bodenbelag', 'boden', 'parkettboden', 'dielen', 'böden', 'parkettböden'	2.0 %
15	food	wein', 'weine', 'trauben', 'weingut', 'reben', 'winzer', 'rebsorten', 'weinbau', 'weinberg'	2.0 %
16	agriculture	boden', 'dünger', 'pflanzen', 'nährstoffe', 'kompost', 'pflanze', 'düngung', 'erde', 'stickstoff', 'wachstum'	2.0 %
17	agriculture	tiere', 'hof', 'stall', 'kühe', 'rinder', 'schweine', 'betrieb', 'hühner', 'haltung', 'eier'	1.9 %
18	agriculture	biogas', 'biogasanlage', 'biomasse', 'biogasanlagen', 'strom', 'anlage', 'energie', 'anlagen', 'wärme', 'energien'	1.9 %
19	agriculture	früchte', 'sorten', 'ernte', 'erdbeeren', 'äpfel', 'sorte', 'obst', 'streuobstwiesen', 'apfel', 'beeren'	1.7 %
20	food	pfeffer', 'salz', 'tomaten', 'gemüse', 'salat', 'knoblauch', 'sauce', 'schneiden', 'olivenöl'	1.6 %
21	wood	wpc', 'holz', 'terrassendielen', 'terrasse', 'lärche', 'garten', 'gartenhaus', 'dielen', 'carport', 'sichtschutz'	1.5 %
22	wood	holz', 'oberfläche', 'holzes', 'pflege', 'außenbereich', 'schutz', 'oberflächen', 'holzschutz', 'öl', 'reinigung'	1.5 %
23	food	nüsse', 'zucker', 'zutaten', 'müsli', 'geschmack', 'lecker', 'einfach', 'kannst', 'mandeln', 'snack'	1.4 %
24	food	bienen', 'honig', 'insekten', 'imker', 'wildbienen', 'imkerei', 'manuka', 'nektar', 'bienenvölker', 'pollen'	1.4 %
25	agriculture	brot', 'getreide', 'mehl', 'weizen', 'dinkel', 'backen', 'backwaren', 'sauerteig', 'roggen', 'mühle'	1. 3%
26	wood	baum', 'bäume', 'baumpflege', 'bäumen', 'baumes', 'fällen', 'äste', 'baumfällung', 'fällung'	1.2%
27	wood	holz', 'fenster', 'türen', 'holzfenster', 'haustüren', 'alu', 'tür', 'kunststoff', 'innentüren'	1.1 %
28	water	fisch', 'lachs', 'fische', 'kaviar', 'forellen', 'fleisch', 'aquakultur', 'aal', 'matjes', 'forelle'	1.1 %
29	water	wasser', 'pflanzen', 'pflanze', 'erde', 'bewässerung', 'wurzeln', 'gießen', 'boden', 'blätter', 'topf'	1.1 %
30	food	kräuter', 'kräutern', 'wildkräuter', 'natur', 'pflanzen', 'küche', 'gewürze', 'gemüse', 'wildpflanzen', 'heilpflanzen'	1.0 %
31	food	käse', 'milch', 'käsesorten', 'molkerei', 'geschmack', 'käserei', 'rohmilch', 'joghurt', 'weichkäse'	0.9 %
32	agriculture	maschinen', 'ernte', 'bodenbearbeitung', 'mais', 'landwirtschaftlichen', 'einsatz', 'mähdrescher', 'technik', 'aussaat', 'landtechnik'	0.9 %

continued on the next page

Table E.1: Topic model based on bioeconomy paragraphs – continued

Topic	Name	Top words	Share
33	food	gin', 'rum', 'whisky', 'geschmack', 'aromen', 'aroma', 'botanicals', 'alkohol', 'vanille', 'destilliert'	0.8 %
34	food	kaffee', 'bohnen', 'kaffees', 'kaffeebohnen', 'arabica', 'espresso', 'robusta', 'fairtrade', 'bohne', 'kakao'	0.8 %
35	water	koi', 'teich', 'aquarium', 'algen', 'fische', 'wasser', 'pflanzen', 'wasserpflanzen', 'gartenteich', 'futter'	0.7 %
36	cosmetics & food supplements	protein', 'aminosäuren', 'körper', 'eiweiß', 'ernährung', 'proteine', 'soja', 'whey', 'veganer', 'vitamin'	0.7 %
37	cosmetics & food supplements	cbd', 'cannabis', 'thc', 'hanf', 'öl', 'cannabidiol', 'hanfpflanze', 'cannabinoide', 'wirkung', 'blüten'	0.7 %
38	agriculture	rasen', 'rollrasen', 'mähen', 'rasenfläche', 'kunstrasen', 'vertikutieren', 'rasens', 'unkraut', 'moos', 'grün'	0.7 %
39	food	tee', 'tees', 'mate', 'geschmack', 'blätter', 'tasse', 'bio', 'matcha', 'aroma', 'grüntee'	0.7 %
40	food	öl', 'fettsäuren', 'omega', 'samen', 'öle', 'kokosöl', 'leinöl', 'rapsöl', 'gewonnen', 'olivenöl'	0.6 %
41	wood	holzspalter', 'stihl', 'forst', 'motorsäge', 'gerät', 'geräte', 'akku', 'motorsägen', 'arbeiten', 'häcksler'	0.6 %
42	agriculture	kartoffeln', 'kartoffel', 'anbau', 'gemüse', 'zwiebeln', 'knollen', 'pommes', 'sorten', 'speisekartoffeln'	0.5 %
43	food	grill', 'grillen', 'bbq', 'holzkohle', 'smoker', 'grillgut', 'fleisch', 'grills', 'temperatur', 'räuchern'	0.5 %
44	wood	bett', 'holz', 'matratze', 'betten', 'massivholz', 'schlafzimmer', 'lattenrost', 'kopfteil', 'liegefläche', 'höhe'	0.4 %
45	construction	dachbegrünung', 'dach', 'begrünung', 'gründach', 'dachbegrünungen', 'extensive', 'dächer', 'begrünte', 'pflanzen', 'vorteile'	0.4 %
46	waste and transport	paletten', 'europaletten', 'transport', 'ippc', 'holzpaletten', 'palette', 'kisten', 'holz', 'ispm', 'holzverpackungen'	0.4 %
47	waste and transport	altholz', 'entsorgung', 'container', 'entsorgen', 'abfälle', 'verwertung', 'holz', 'grünschnitt'	0.4 %
48	wood	friedwald', 'urne', 'verstorbenen', 'asche', 'baumbestattung', 'baum', 'beigesetzt', 'ruheforst', 'beisetzung', 'urnen'	0.4 %
49	food	pilze', 'pilz', 'vitalpilze', 'pilzen', 'shiitake', 'pulver', 'vitalpilz', 'reishi', 'vitalpilzen', 'champignons'	0.3 %
50	wood	weihnachtsbaum', 'weihnachtsbäume', 'baum', 'nadeln', 'tanne', 'nordmanntanne', 'weihnachten', 'weihnachtsbäumen', 'tannenbaum', 'christbaum'	0.3 %
51	construction	kork', 'korkboden', 'korkböden', 'rinde', 'korkeiche', 'bodenbelag', 'elastisch', 'boden', 'eigenschaften', 'linoleum'	0.3 %
52	wood	treppe', 'treppen', 'holztreppen', 'holztreppe', 'stufen', 'handlauf', 'buche', 'geländer', 'holz', 'eiche'	0.3 %
53	food	bier', 'hopfen', 'malz', 'hefe', 'brauerei', 'biere', 'würze', 'brauen', 'gerste', 'gebraut'	0.3 %
54	wood	klang', 'instrument', 'hölzer', 'instrumente', 'ahorn', 'mahagoni', 'holz', 'ebenholz', 'korpus', 'gefertigt'	0.3 %
55	wood	sauna', 'saunen', 'saunaofen', 'holz', 'mm', 'harvia', 'espe', 'finnischen', 'fichte', 'finnische'	0.2 %

Online Appendix

The online appendix is available at https://osf.io/yvfwh/.