

# Machine learning and the identification of Smart Specialisation thematic networks in Arctic Scandinavia

Moilanen Mikko, Østbye Stein & Simonen Jaakko

To cite this article: Moilanen Mikko, Østbye Stein & Simonen Jaakko (2021): Machine learning and the identification of Smart Specialisation thematic networks in Arctic Scandinavia, *Regional Studies*, DOI: [10.1080/00343404.2021.1925237](https://doi.org/10.1080/00343404.2021.1925237)

To link to this article: <https://doi.org/10.1080/00343404.2021.1925237>



© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 16 Jun 2021.



[Submit your article to this journal](#)



Article views: 921



[View related articles](#)



[View Crossmark data](#)

# Machine learning and the identification of Smart Specialisation thematic networks in Arctic Scandinavia

Moilanen Mikko<sup>a</sup> , Østbye Stein<sup>b</sup>  and Simonen Jaakko<sup>c</sup> 

## ABSTRACT

The European Union (EU) has recognized that universities and research institutes play a critical role in regional Smart Specialisation processes. Our research aims to identify thematic cross-border research domains across space and disciplines in Arctic Scandinavia. We identify potential domains using an unsupervised machine-learning technique (topic modelling). We uncover latent topics based on similarities in the vocabulary of research papers. The proposed methodology can be utilized to identify common research domains across regions and disciplines in almost real time, thereby acting as a decision support system to facilitate cooperation among knowledge producers.

## KEYWORDS

Smart Specialisation; cross-border cooperation; innovation; machine learning

JEL O31, R11, R58

HISTORY Received 6 May 2020; in revised form 6 April 2021

## INTRODUCTION

Smart Specialisation policy emphasizes an entrepreneurial-discovery process where the identification of domains of advantage should emerge through a multi-actor process involving local agents rather than through top-down, centralized bureaucratic processes of technology choice and selection (Foray, 2014). Immediately, such an approach raises three practical problems: What does 'local' refer to? How should the relevant agents and domains of advantage be selected? The domain of advantage is not necessarily contained within the borders of existing administrative regions, and the relevant local agents may well come from different regions or even from different sides of national borders (Muller et al., 2017). Another challenge is helping agents discover domains of advantage (Capello & Kroll, 2016).

On the basis of these introductory remarks on Smart Specialisation, we would like to suggest the following two issues for closer scrutiny:


- The idea of domains transcending administrative borders suggests the need to learn more about Smart Specialisation in cross-border regional innovation systems (RISs).

- The idea of multi-actor involvement in identifying relevant research domains across borders indicates the need to learn more about how such identification could be accomplished in practice.


In this paper we address these two issues by looking at the domains of knowledge and innovation in the northern periphery of Europe – more specifically, Arctic Norway, Sweden and Finland (Arctic Scandinavia). We focus on the so-called Arctic 5 university cities: Luleå and Umeå in Sweden, Oulu and Rovaniemi in Finland, and Tromsø in Norway (see Appendix A in the supplemental data online for more information).

As a starting point, we acknowledge the possibility that a Smart Specialisation policy could improve cross-border cooperation in innovation among the Arctic 5 cities and thereby improve innovation capacities. However, making this kind of strategy work requires agents to have a good understanding of the strengths in research and innovation (R&I) of not only their home region but also other regions. Moreover, agents cooperating across regions and borders need to share a common vision to know where to place their focus.


## CONTACT

<sup>a</sup> (Corresponding author)  mikko.moilanen@uit.no


School of Business and Economics, UiT The Arctic University of Norway, Tromsø, Norway.

<sup>b</sup>  stein.ostbye@uit.no

School of Business and Economics, UiT The Arctic University of Norway, Tromsø, Norway.

<sup>c</sup>  jaakko.simonen@oulu.fi

Department of Economics, Accounting and Finance, Oulu Business School, University of Oulu, Oulu, Finland.

 Supplemental data for this article can be accessed at <https://doi.org/10.1080/00343404.2021.1925237>.

Lundquist and Trippel (2013) persuasively argue that the development of cross-border innovation systems should be viewed as a process involving different stages depending on how well integrated the innovation systems are. A recent report on cross-border cooperation in innovation in the Scandinavian Arctic showed that this cross-border innovation system is weakly integrated (University of Oulu, 2019). On this basis, we aim to identify possible shared visions among agents within and across borders as a first step towards a semi-integrated system (Muller et al., 2017).

In Smart Specialisation jargon, we might say with slightly more precision that we are looking for a methodology that can be useful for identifying a shared vision of research domains. In poorly integrated economies, there may be no meaningful conversations about what a shared vision could look like. We therefore recognize the potential for improving outcomes by enhancing the market mechanism through the provision of additional information to agents to mitigate communication and coordination failures.

The European Commission has suggested that in the design phase of Smart Specialisation policy, universities and research institutes have an important role to play in identifying research domains with significant strengths and high potential at the national or regional level, as well as in assisting regions to look outside their boundaries to compensate for limited local capacity to absorb research output (European Commission, 2014).

As a starting point, we use the common ground theory of cognitive interdisciplinarity, which states that a common vocabulary is an important condition for knowledge-sharing and collaboration among researchers (Bromme, 2000). At a very basic level, as a prerequisite for cooperation, we are trying to identify common research domains across the Arctic 5 universities and research institutes based on a common terminology (Huber, 2012). This can provide some guidance for agents by focusing on themes that are potentially meaningful for conversations across borders. Ultimately, this may help the search for common ground and for fruitful focus areas for cross-border cooperation in innovation, consistent with the underlying idea behind Smart Specialisation.

In the present paper we specifically examine the potential for using machine-learning tools in this identification process. Based on a sample of 10,000 recent research paper abstracts from the Arctic 5 universities, we use a topic modelling algorithm that identifies 26 potential shared visions (latent topics) that could be interpreted as candidates for possible shared research domains or even domains of Smart Specialisation. Furthermore, we use this method to detect correlations between topics to establish which topics are shared by universities and the strength of these relationships. We also offer an analysis of the shared topics and how they connect to shed light on the potential for cross-fertilization across disciplines and regions, which may advance innovation.

The remainder of the paper is structured as follows. We next discuss the concept of Smart Specialisation in

more detail and then the importance of proximity for innovation. We then introduce the data and research methods in the third section. The fourth section presents the results. The final section provides the main conclusions of the analysis.

## THEORETICAL VIEWS ON CROSS-BORDER INNOVATION COOPERATION

### Smart Specialisation Strategy as a tool for cross-border innovation cooperation

The European Commission launched the Lisbon Agenda in 2000 with the goal of making Europe 'the most competitive and dynamic knowledge-based economy in the world' by 2010 (e.g., van Ark et al., 2008). Over time and as 2010 approached, it became clear that this would not happen. In response, the Lisbon Agenda was replaced by the equally ambitious European Union (EU) 2020 growth agenda. This agenda was based on the strategy of Smart Specialisation, a novel concept that was gradually developed and made popular as a new basis for what the Lisbon Agenda had promised but failed to deliver: increased innovation as an engine of growth aiming to make the EU the most competitive knowledge economy in the world (e.g., Balland et al., 2019). The Smart Specialisation Strategy equally aims to support EU regional policy and cohesion (McCann & Ortega-Argilés, 2015).

What are the key principles of a strategy for Smart Specialisation? Smart Specialisation is a place-based approach. In other words, it builds on the assets and resources available to regions and on their specific socio-economic challenges, and the idea is to identify areas for future growth. Smart indicates that regions should be able to identify their knowledge-based assets. Specialization suggests that regions should prioritize their R&I investments in areas where they are competitive. A strategy, in turn, is a shared vision defined by regional stakeholders for the long-term development of regional innovation (McCann & Ortega-Argilés, 2014).

The architects behind the Smart Specialisation Strategy claim that this new strategy is not another example of the controversial 'picking winners' strategy. In line with this claim, they insist that the identification of domains of advantage should emerge through a multi-actor process involving local agents. One challenge is that it is not completely clear who these agents should be; another is that the domains of advantage are not necessarily contained within the borders of existing administrative regions, implying that the relevant agents may well come from different regions or even from different sides of national borders. Furthermore, it would be problematic if those assigned such roles emphasize domains that may be considered strategic at the national and global levels but lack local foundations (Capello & Kroll, 2016).

According to Muller et al. (2017), regions often cannot pursue, and do not have to pursue, everything in terms of science, technology and innovation on their own. Rather,

they can specialize in carefully selected specific domains and try to find synergistic advantages by interacting with other regions. This is particularly true for Arctic regions, which have limited resources. All regions have certain economic, technological and knowledge-based assets and strengths that can be utilized effectively to bring about growth and economic transformation (Foray, 2013). However, to realize their full potential, it may be necessary to look for complementarities across regions.

The selected set of priorities should focus on the existing strengths of the regional economy and the emerging opportunities within and across regions. The selection process itself must be based on versatile qualitative and quantitative information on the different areas of expertise in the regions. Regional stakeholders – for instance, research groups in universities and research centres in different fields of expertise – could play an important role in identifying not only promising areas of regional specialization but also weaknesses that are currently hampering innovation (Foray et al., 2009).

Therefore, for a Smart Specialisation-type policy to work in a regional context, the analytical focus must centre on ways to maximize knowledge spillovers and learning linkages within the region and between regions. This can be promoted by identifying thematic areas that might connect the research domains of the different Arctic 5 cities. Such an approach could be especially beneficial for cities that lack complementary domains, as it would identify if those themes and technologies are available in another city.

Following the arguments for Smart Specialisation, in the case of interregional cooperation, regional policy should focus on the most connected industries in peripheral regions so that the regional industrial base is best able to learn from more advanced regions. Less advanced regions might capture knowledge spillovers from leaders, and leading regions may receive ideas from less advanced regions that help them reinvent themselves (Foray, 2013). Selected areas of Smart Specialisation are typically areas of expertise that are at the intersection of different sectors, technologies or knowledge domains. The identification of the main thematic areas of research and their connections to one another provides a good basis for facilitating cross-border cooperation and a larger, more systematic view of the possibilities of Smart Specialisation.

The aim of our research is to investigate, through the lens of Smart Specialisation, whether Arctic 5 cities have competence areas that might provide new possibilities for intensified cross-border cooperation involving representatives of both industry and academia. What are the thematic research areas where these cities have the potential to generate innovation activities to support knowledge-driven growth?

According to Muller et al. (2017), cross-border cooperation among universities and research centres can enhance the integration of RISs. One step in this direction is to provide information to the regions to learn whether there is the potential to find and exchange ideas across both research fields and borders. Novel combinations of

existing knowledge require that the fields of research that cross borders are sufficiently similar.

Both Bromme (2000) and Huber (2012) emphasize the need for a common vocabulary, a subdimension of cognitive proximity, as a prerequisite for the functioning of knowledge networks. In our research, we apply this same idea and analyse cross-border regions' shared research domains. Recognition of this type of cognitive proximity will significantly improve cross-border cooperation in innovation and can help to find common fields for Smart Specialisation.

### Challenges to the cross-border innovation system

Megatrends such as globalization, digitalization, the growing role of the service sector (especially in the Western world), urbanization and the agglomeration of economic activities, and the ageing of the population also affect development in the Scandinavian Arctic regions. These megatrends lead to challenges and threats as well as opportunities. Broad and systematic cross-border cooperation in various fields of business and innovation involving industry and academia could significantly reinforce the regions' ability to tackle these issues.

Why has it proved difficult to find a common vision for cross-border cooperation in innovation? First, differences in national innovation policies and Smart Specialisation strategies have an impact. Political visions and strategic priorities vary from country to country, and this can also be seen among the Scandinavian Arctic regions (Kristensen et al., 2018). Furthermore, the priorities of the Arctic strategies in Finland, Norway and Sweden seem to vary quite significantly (Karlsdóttir & Greve Harbo, 2017).<sup>1</sup>

Traditional proximity issues (geographical, technological, cultural, cognitive, etc.) clearly also play an important role in this context (e.g., Lundquist & Trippl, 2013). In the case of the Arctic 5 cities, the geographical distance between the cities is great because they are located 350 km apart, on average (Figure 1). However, as Makkonen et al. (2017) note, geographical distance does not preclude the efficient diffusion of technological knowledge. Cultural, cognitive, institutional and social proximity as well as similarities in technical expertise can facilitate knowledge diffusion and the integration process across regions despite geographical distance (Boschma & Frenken, 2009; Muller et al., 2017).

Furthermore, as already discussed, the innovation policy and innovation activities of cross-border regions may be coordinated based on different national innovation priorities. Guidelines for national innovation policies and targets for cross-border cooperation in innovation may not be aligned. Such misalignment could reduce firms' interest in initiatives promoting cross-border cooperation in innovation (Lundquist & Trippl, 2013), and under limited regional political autonomy, in particular, it can be difficult to implement cross-border Smart Specialisation strategies (Makkonen et al., 2017).

Why try to increase cooperation in innovation among the Arctic 5 cities? One reason is that they are all located



**Figure 1.** The Arctic 5 cities in northern Finland, Norway and Sweden.

in the Northern Sparsely Populated Areas (NSPA), and they all face more or less the same challenges and opportunities presented by the Arctic environment. This creates a good starting point for innovation cooperation. This kind of innovation cooperation would also be in line with Organisation for Economic Co-operation and Development (OECD) policy recommendations, which encourage NSPA to collaborate on joint opportunities related to their Smart Specialisation strategies. There is also genuine political interest in increasing cross-border cooperation in innovation in the north.<sup>2</sup> The goal of this cooperation would be to combine the expertise held by these Arctic 5 universities and cities in different fields and to share their plans and ideas for the future.<sup>3</sup>

Following Lundquist and Trippel (2013), we can argue for the existence of mutual understanding and trust and a shared organizational culture among the Arctic 5 cities, which provides solid grounds for knowledge exchange

and collaboration. The current state of the RIS formed by these cities can be considered weakly integrated, characterized by, for example, low levels of knowledge interaction and innovation linkages in only a few selected fields (Lundquist & Trippel, 2013; Makkonen et al., 2017). However, there is great potential for a more integrated cross-border innovation system (University of Oulu, 2019).

University cities are in many ways the engines of growth in their regions, and they have a positive economic impact on the hinterland. Their various R&I activities drive knowledge transfer between not only universities but also peripheral regions and university cities. Multidisciplinary universities play a leading role in innovation activities that is becoming increasingly important in a world where most innovations take place at the intersections of different branches of science. Interdisciplinary collaboration is becoming an increasingly integral part of

research, and its importance has been emphasized in EU R&I programmes. As key knowledge producers, universities and other research institutions play an important role in designing and implementing Smart Specialisation strategies. Furthermore, international networks of universities and research institutes offer regions the opportunity to find solutions beyond their national borders (European Commission, 2014).

The Arctic 5 universities have globally recognized research facilities at their disposal and expertise in several fields, for example, health, biotechnology, various fields of engineering, and information and communication technology (ICT). Furthermore, ongoing cooperation in research in various fields is already occurring among the universities. To date, however, this collaboration has mainly been project based and reliant on personal networks. In other words, there is room for more intense and systematic cooperation across borders among the actors of this regional innovation ecosystem.<sup>4</sup>

Another good starting point for intensified innovation cooperation among the Arctic 5 cities is that each of these cities has its own strong innovation sector, which provides a good basis for knowledge flows across cities. In other words, the functional distance between regions is not that great. This means that these cities all have the opportunity to learn something new from each other through more intensified research cooperation. The concept of 'joint specialisation' presented by Muller et al. (2017) describes well the potential for the Arctic 5 cities to develop cooperation in innovation.

Huber (2012) concludes that similarity in terms of vocabulary is the most critical form of cognitive proximity (indeed, even a prerequisite for cognitive proximity) and is vital for effective communication. Our empirical analysis applies this idea of 'similarity in language' to identify thematic topics of research and the linkages between them both within and between among the Arctic 5 cities.

## DATA AND RESEARCH METHODS

Machine-learning and real-time information processing can be used to support the knowledge-discovery process within Smart Specialisation. In this paper we employ an unsupervised machine-learning text-processing technique called topic modelling. We study similarities in the vocabulary that researchers use to identify common research domains as the basis for distinguishing the domains of advantage that may ultimately foster collaboration among the Arctic 5 cities.

The framework we employ, structural topic modelling (STM) (Roberts et al., 2019), processes a very large set of research documents, detects word patterns within them, and automatically identifies shared latent research domains quickly and reproducibly. Because STM does not rely heavily on the prior assumptions of the researcher, it is very useful for our exploratory study, the aim of which is to establish an overall idea of the topics being addressed by the research conducted in the Arctic 5 cities.

Topic modelling has increasingly been used in the context of mapping regional assets and advantages. For instance, Papagiannidis et al. (2018) employ it to detect industrial clusters, using information from companies' webpages. Pavone et al. (2019) use text from the Eye@RIS3 platform to classify regional priorities across the EU-28. As far as we are aware, our study using research paper abstracts is the first to study the potential of topic modelling for assisting scientific collaboration by identifying common research domains among universities.

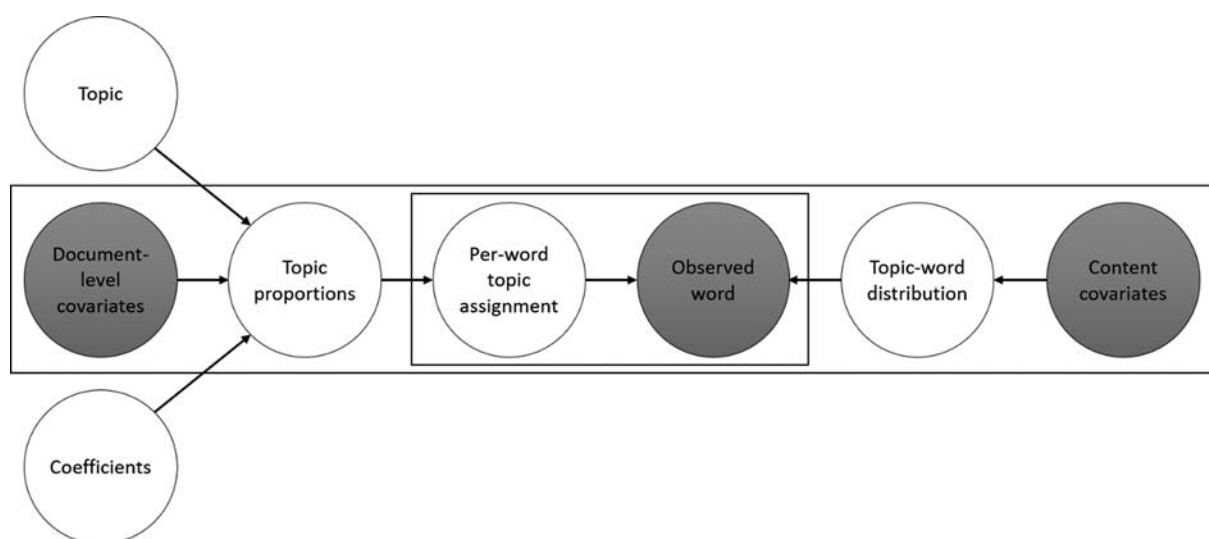
### Data retrieval and reprocessing

Our initial data set consists of 10,000 abstracts. We gathered 2000 abstracts from the most recent scientific papers for which at least one of the authors is affiliated with a university or research institute in an Arctic 5 city. This was done for each of the five Arctic 5 cities and resulted in a total of 10,000 abstracts. The abstracts were retrieved from Scopus through to 6 January 2020.<sup>5</sup> Of the abstracts collected, 433 lacking proper abstract information were excluded. We extracted *only nouns* from the abstracts for use in our analysis, resulting in 9577 analysable abstracts for the topic modelling analysis. Our goal was to discover latent topics in the articles using the abstracts, to uncover correlations between topics and to estimate their relationship to the researchers' affiliated cities. To reduce the computational load, we used nouns that appeared at least 50 times in our structured set of all analysable abstracts (called the corpus in linguistics). Before conducting the topic modelling analysis, we also improved the quality of the data by removing nouns that are very commonly used in abstracts (such as approach, study, model, etc.).

### Method

STM is a probabilistic topic modelling method that integrates machine learning with causal inference mechanisms in a generalized framework. Topic coverage and word distribution are approximated with Bayesian inference (Roberts et al., 2016, 2019). STM is an extension of probabilistic topic models such as *latent Dirichlet allocation* (Blei et al., 2003) and *correlated topic models* (Blei & Lafferty, 2007). In STM, documents (abstracts in our case) represent an unknown mixture of latent topics, meaning that a single abstract is composed of multiple topics. A topic is a probability vector over the words in the vocabulary (the vocabulary in our analysis consists of the nouns in the abstracts). All words are therefore potentially present in all topics, albeit with different weights.

Unlike earlier topic models, STM allows the connection of metadata. Using STM, we are able to connect documents to information about the authors' affiliation cities. We can link this information to the degree of association between a document and a topic (topic prevalence) as well as to the degree of association between a word and a topic (content prevalence). This allows us to compare the probabilities that a topic occurs in the abstracts from the different Arctic 5 cities. The model is illustrated in Figure 2 (see Appendix B in the supplemental data online for more details).



**Figure 2.** The structural topic model.

Note: Symbols and notation are explained in more detail in Appendix A in the supplemental data online.

Source: Roberts et al. (2016).

The model identifies the topic proportions in an abstract using information from the vector of covariates. STM also explicitly models the interdependence of topics in terms of topic correlations. This allows the latent factors to be described by a combination of topics that are closer together in higher order dimensions of the central latent concept.

## RESULTS

### Number of topics

The number of topics that are related to the corpus of abstracts is not known and must be estimated. This can be done in various ways. In this study we use the metrics proposed by Deveaud et al. (2014), who utilize a measure that maximizes the divergence across topics. We tested a range of topic numbers from two to 100 and ended up with 26 topics. To interpret and name the topics, we must understand their thematic meanings. We focus on the most frequent words and on those that are important in distinguishing between topics (Bischof & Airoldi, 2012; Roberts et al., 2019). We report this process and the representative words in Table B1 in Appendix B in the supplemental data online.

Although this method organizes latent domains across large amounts of text, the utility of these domains ultimately depends upon the thoughtful and subjective assignment of meaning to the domains, as current topic modelling techniques require the manual labelling of topics. Automated content analysis methods, such as topic modelling, are no substitute for careful thought and reasoning (Grimmer & Stewart, 2013). Although we use different metrics to examine the validity of the identified topics, topic identification still requires manual coding and interpretation. Therefore, validity must be ensured when applying these methods. In the absence of standard validation procedures, we choose to follow the

suggestions of DiMaggio et al. (2013) and test the model's statistical and semantic validity (for details on these tests, see Appendix C in the supplemental data online). Our test results show that the statistical validity is good (high exclusivity of topics and good semantic coherence). Our tests for semantic validity (how well the topics correspond to groupings that are natural for humans and how the mixture of topics agrees with human associations) are on a par with the models in Chang et al. (2009) and Arnold et al. (2015).

### Topic proportions

In addition to identifying research topics, we investigate the prevalence of each topic by calculating so-called topic proportions. A topic is a mixture of words where each word has a specific probability of belonging to that topic. An abstract, in turn, is a mixture of topics, meaning that a single abstract can be composed of multiple topics. The sum of the topic proportions across all topics for a document is thus equal to 1. To indicate how large each of the 26 topics is in our corpus, we calculate expected topic proportions. The results are shown in Figure 3. *Business and innovation* are the most common topics, while *haematology* is the least common.

We also obtain 26-word probability vectors over the vocabulary for each of the Arctic 5 cities. Figure 4 shows the topic proportion distributions by Arctic 5 city. These distributions show how the different cities contribute to each individual topic, illustrating how they are focusing their research on specific topics of interest, reflecting research priorities and traditions. For example, Luleå, with its university of technology, has a strong focus on technology. Cities that have university hospitals (Oulu, Tromsø, Umeå) are strongly represented in medicine and health. Figure 4 also shows how diversified the research of a region can be from a thematic perspective.

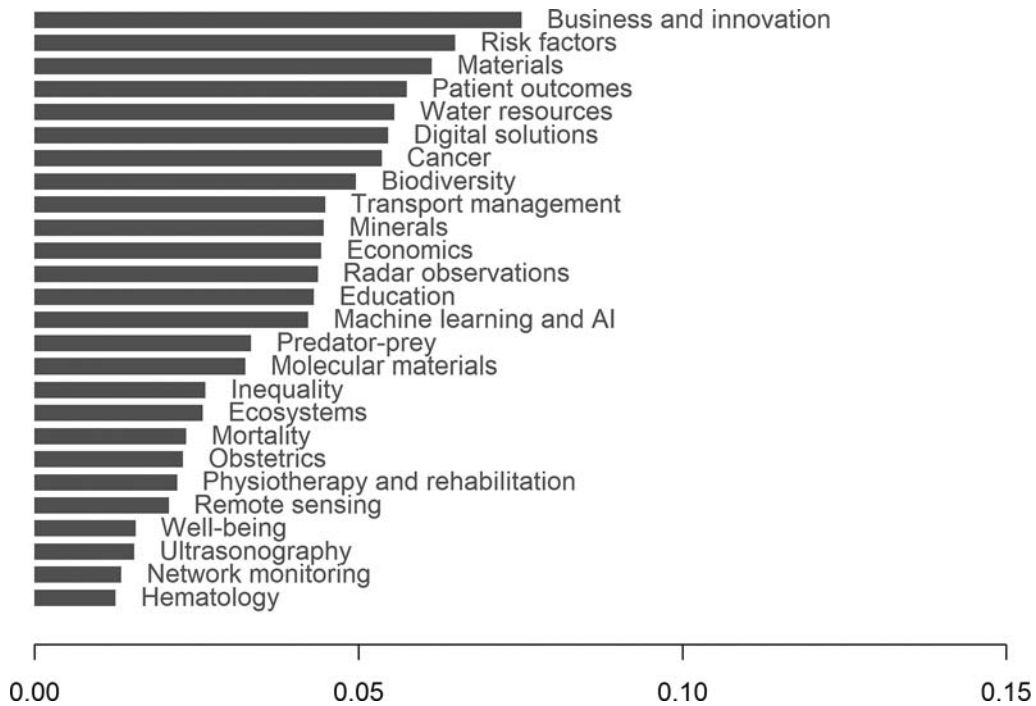


Figure 3. Topic labels and expected topic proportions in the corpus.

For example, the research themes in Tromsø and Umeå are clearly more diversified than those in Luleå.

We can also use STM to discover how topics are correlated by analysing how often topics co-occur within the same abstract. The interpretation of these connections is that the presence of correlation represents proximate research topics that have a greater potential for influencing one another. Figure 5 shows the correlations between topics; a shorter distance between topics means a stronger correlation (correlations < 0.05 are not shown). The sizes

of the nodes are scaled according to their topic proportions in the corpus: the larger is the node, the more prevalent that topic in our corpus. Figure 5 also shows clusters of topics that are densely connected within the cluster but sparsely connected to other clusters. We use edge betweenness cluster structure detection, which is based on the idea that it is likely that edges that connect separate clusters have a high level of edge betweenness, as all the shortest paths from one cluster to another must traverse through them (Newman & Girvan, 2004). Using this

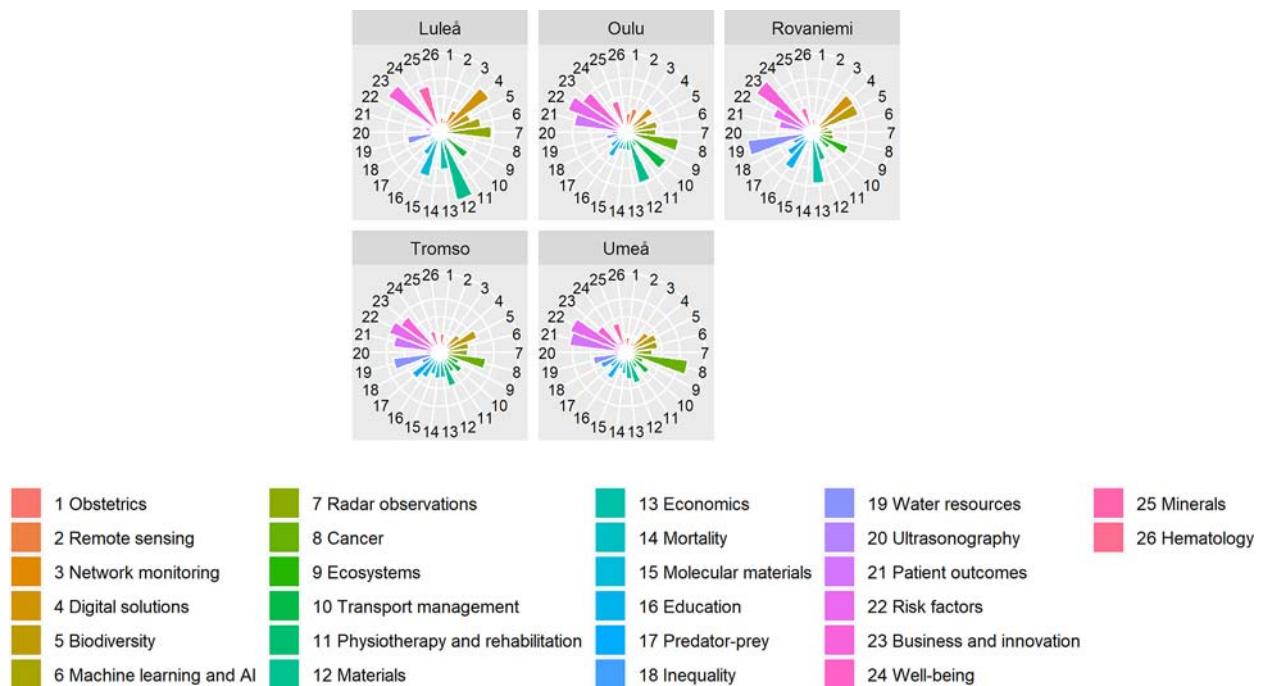
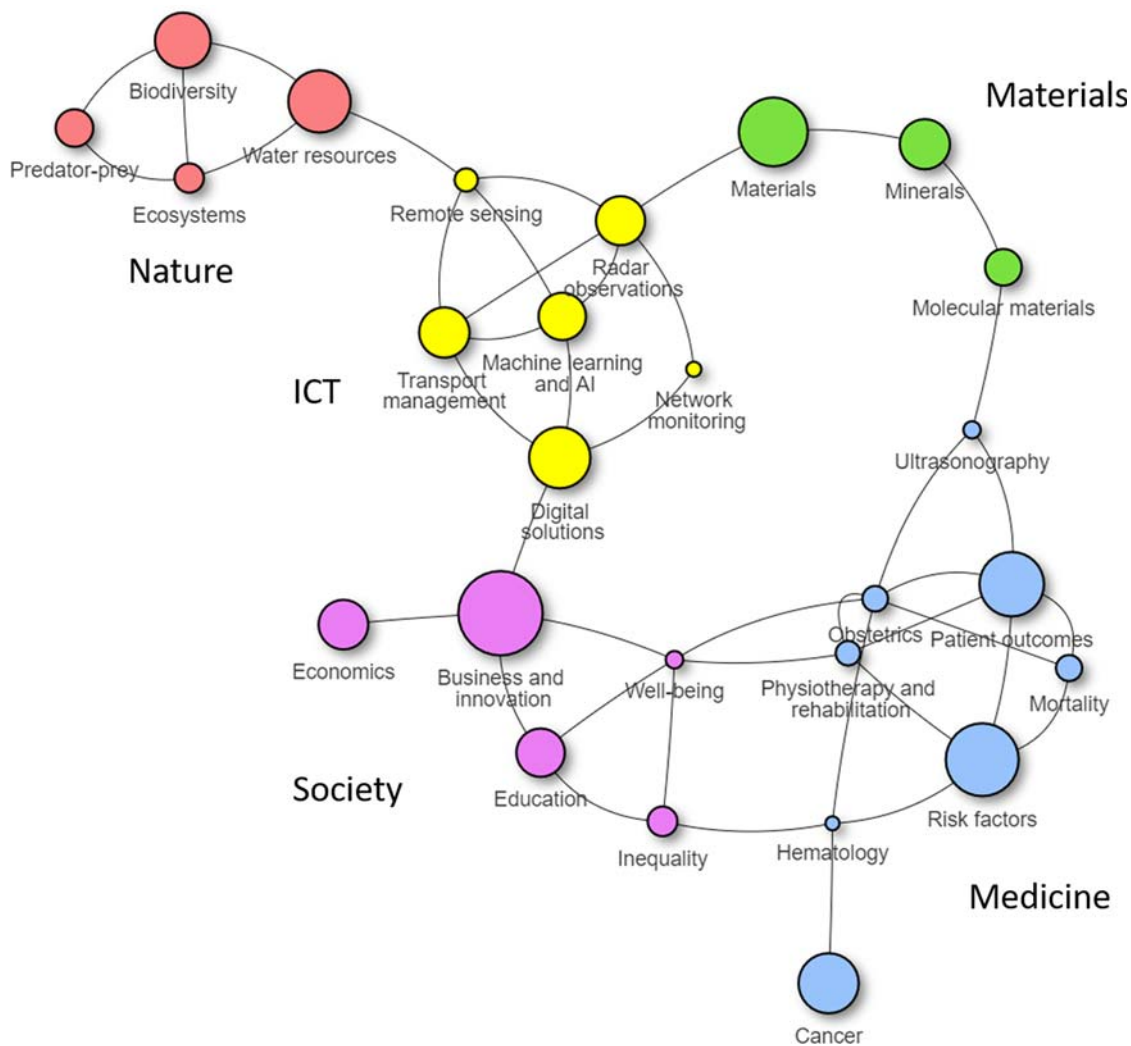


Figure 4. Distributions of topic proportions for the different Arctic 5 cities. The proportions sum to unity within a city.





**Figure 5.** Thematic network of the research in the Arctic 5 cities. The nodes are scaled according to their topic proportions in the corpus.

idea, the algorithm identifies five clusters: nature, ICT, materials, medicine and society.

### Identification of shared domains

We now turn our attention to exploring the research domains to identify potential domains for cooperation. A high level of similarity in vocabulary is a prerequisite for effective communication and therefore for potential cooperation. To measure the overlap in vocabulary, we compute the pairwise cosine similarity (CS),<sup>6</sup> which is often used to measure document similarity in text analysis. We calculate the CS between the vocabulary distributions for pairs of Arctic 5 cities. A higher CS indicates a large overlap in vocabulary, meaning that researchers from the two cities use a ‘shared vocabulary’.

As a first step, we look at the vocabulary similarities when writing about the same topic. Figure 6 shows the similarity in language within a domain across the Arctic 5 cities. We also calculated the mean CS to measure a topic’s overall level of similarity (in parentheses). The heatmaps show that some medicine-related themes have high similarity in their technical languages, while *remote*

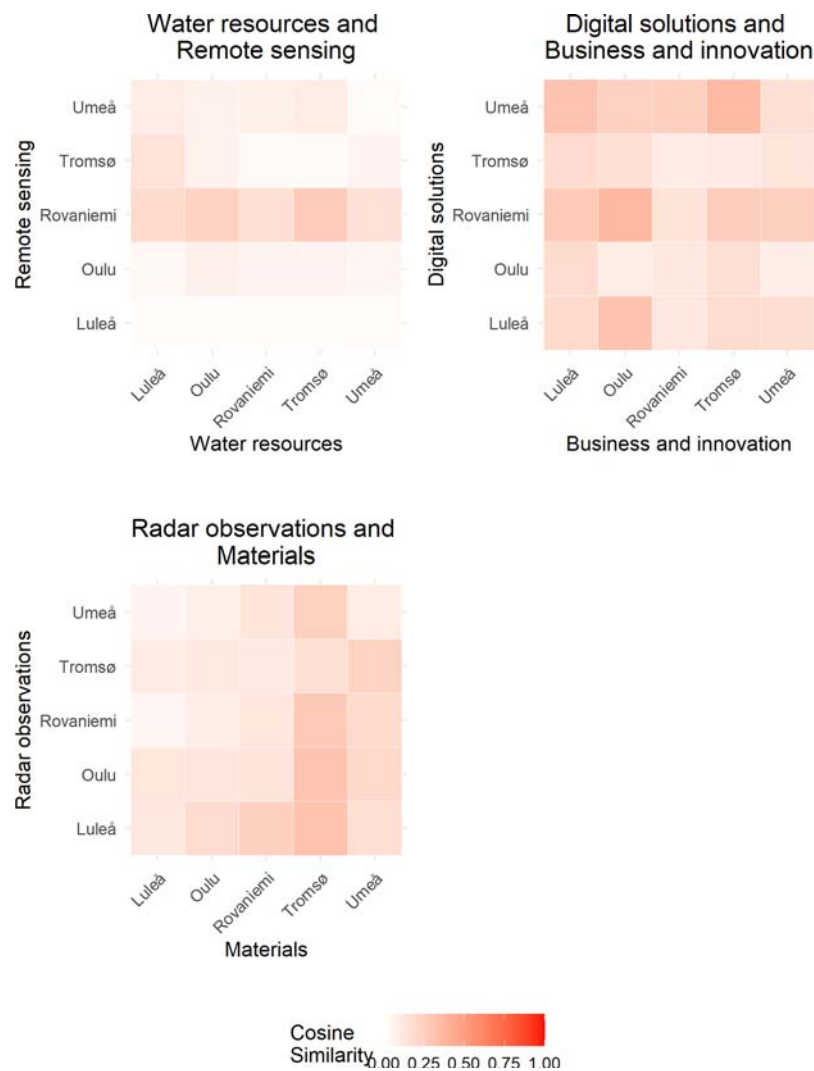
*sensing* and *network monitoring* have the lowest similarities. This can be interpreted as indicating that similarities in vocabulary are highest among the various topics in medicine and that potential exists for cooperation on topics within medicine between cities. Likewise, the heatmap for the *business and innovation* topic reveals possibilities for cooperation within that topic (high mean CS = 0.72).

Our results indicate where bilateral cooperation could be worth the effort. For example, in the case of *remote sensing*, the overall similarity in language is low, but the vocabularies in Tromsø and Luleå seem to be closest to each other.

Our model sheds light on vocabulary similarities within a given topic but also on the shared vocabulary across the Arctic 5 universities between any topics of interest: we are able to identify ‘bridges’ that bind together thematic clusters. It is often suggested that great advances in R&I take place at the intersections between disciplines. Look at an example of how the bridges between domain clusters are built. Although the similarity of language is not high for *remote sensing*, it plays an important role in bridging the ICT and nature clusters (Figure 5). The



**Figure 6.** Heatmaps of pairwise cosine similarities between Arctic 5 cities by topic. Note: Mean cosine similarity (CS) is shown in parentheses.



**Figure 7.** Heatmaps of cosine similarities between topics bridging thematic clusters by Arctic 5 city.

heatmap in the north-west corner of Figure 7, for instance, shows that the vocabulary in the *remote sensing* topic in Rovaniemi has some similarities with the vocabularies of the *water resources* theme in other universities.

By investigating these bridges in more detail, we can learn how to build new bridges between clusters. We may argue that in the case of *business and innovation* and *digital solutions*, the bridge is rather ‘strong’ because of the high similarity in language used across all regions. Furthermore, we can use these maps to provide informed guidance to universities regarding where to look for cooperation. For example, the diagonal in the *business and innovation* and *digital solutions* heatmap in Figure 7 shows that there are very low similarities between these topics *within* each Arctic 5 city. This indicates that there is little common language between *business and innovation* and *digital solutions* research within each Arctic 5 city. However, the same map shows that there is high similarity between the *digital solutions* vocabulary in Luleå and the *business and innovation* vocabulary in Oulu, suggesting a potential for collaboration between these two cities.

Our framework is equally suitable for uncovering similarities in vocabulary among Arctic 5 cities within topics in a cluster. However, following that track would require a paper in itself and thus is beyond the scope of this work. For now, we simply emphasize the vast potential of this framework. To further substantiate this claim, we provide an example in Appendix D in the supplemental data online of a broader application of our framework using the six topics in the ICT cluster to construct an information density heatmap that is presented in Figure D1 online.

Knowledge about similarities in vocabulary and thematic clusters based on a systematic and validated approach along the lines we have suggested could improve matchmaking relative to the current situation, where cooperative projects among the Arctic 5 universities and research institutes occur more or less by chance when some researchers meet each other at seminars and conferences. Furthermore, given information on the city and university affiliations of researchers interested in particular topics, decision-makers could even proactively and directly encourage such interactions.

## CONCLUSIONS

In this paper we addressed Smart Specialisation in the Scandinavian Arctic. It is widely accepted that cross-border cooperation in innovation and networking can improve innovation capacity in sparsely populated regions (e.g., Makkonen et al., 2017). This kind of innovation cooperation is also in line with OECD's policy recommendations,<sup>7</sup> which encourage NSPA to collaborate on joint opportunities related to their Smart Specialisation strategies to compensate for geographical disadvantages.

Differences in national innovation strategies play an important role in efforts to develop cross-border cooperation in innovation. Political visions and strategic priorities vary from country to country, making it difficult to find a common vision. This has clearly been seen in the case of the NSPA (Kristensen et al., 2018). However, mutual understanding and trust and a shared organizational culture provide a good basis for knowledge exchange and research collaboration among the Arctic 5 cities: Oulu, Rovaniemi (Finland), Umeå, Luleå (Sweden) and Tromsø (Norway). These cities have strong academic research, which provides a good starting point for sharing new ideas and for knowledge-based innovation. The EU has also recognized that universities and research institutes play a critical role in regional Smart Specialisation processes.

In this paper we have identified common research domains across the Arctic 5 universities and research institutes to provide guidance for agents, helping them to focus on themes that might be potentially meaningful for conversations across borders and to find areas for cooperation. We follow the idea that a shared vocabulary is an important condition for knowledge-sharing and collaboration between researchers, in line with the common ground theory of cognitive interdisciplinarity. In this study, this was done by analysing the research publications from the Arctic 5 universities and research institutes.

We use topic modelling, a machine-learning text-mining method, to capture the essential role of a shared language in effective communication. Based on a sample of 10,000 recent research paper abstracts from the Arctic 5 universities, we identify potential shared topics. Furthermore, we analyse how the topics connect across the Arctic 5 cities to shed light on the potential for research collaborations and on possible cross-fertilization across disciplines and cities that may advance innovation performance. The intention is for these results to serve as an intermediate input to further analyses within a wider process directed towards a more integrated cross-border innovation system in the Scandinavian Arctic.

The proposed methodology can be utilized to identify common research domains across regions and disciplines in almost real time, thereby acting as a decision support system facilitating cooperation between knowledge producers. The identification of shared domains would not only allow academic collaborations to be better organized but also be beneficial for the entrepreneurial discovery process

and for shaping different innovation policies. It seems clear that text-mining approaches can provide new information and are worth exploring in the context of Smart Specialisation.

It is a long journey from academic work to marketable products, requiring matches with market development, skills and many other competencies. However, the analysis of academic research can help to identify potential promising areas for cooperation in firm-level innovation and to support more fine-grained priority-setting and policymaking. The limited resources of regions require the more precise identification of focus areas and potential for Smart Specialisation not only within the regions but also between the regions. Typical broadly and loosely defined thematic network cooperation may not be the right strategy for such areas; instead, we should concentrate on more precisely identifying those technology areas that will most effectively foster regional growth and provide new innovation opportunities. Research groups of universities and research centres in different fields of expertise have an important role in this search process. It is essential to integrate the analysis presented here with qualitative, participatory and expert-based methods. It is clear that regional stakeholders, especially companies from various industries, must be involved in this process. The successful identification of potential areas of innovation can also provide new opportunities for start-ups.

Topic modelling is a promising approach worth exploring further in the context of Smart Specialisation, as large corpora of research paper abstracts are becoming increasingly accessible. To date, large-scale analyses of research papers with metadata have only been possible using commercial databases such as Scopus. Fortunately, research has moved towards more open-access and open-data initiatives, likely making these kinds of data more openly available in the near future. In addition, the data are available and are being updated continuously, not only for the Scandinavian Arctic but also for other regions in Europe and elsewhere. Hence, this approach can be readily applied outside the geographical area and time frame examined here. In future we envisage that research along the lines suggested here can contribute to better informed policies and research-based guidelines for the implementation of Smart Specialisation in general and cross-border Smart Specialisation in particular.

## DISCLOSURE STATEMENT

No potential conflict of interest was reported by the authors.

## FUNDING

This research is connected to the GenZ project, a strategic profiling project in human sciences at the University of

Oulu. The project is supported by the Academy of Finland [grant number Profi4 318930] and the University of Oulu.

## NOTES

1. There are few initiatives for creating cross-border innovation systems. Some exceptions are the Cross-Border Smart Specialisation Strategy of Galicia–Northern Portugal (Oliveira, 2015) and of the Upper Rhine area, including Alsace in France and Baden-Württemberg in Germany (Muller et al., 2017).
2. For example, see OECD: <http://www.nspa-network.eu/news/oecd-report-launched-in-brussels.aspx>; and an expert group established by the prime ministers of Norway, Sweden and Finland: <https://site.uit.no/growthfromthenorth/files/2015/01/Growth-from-the-North-lowres-EN.pdf>.
3. There are already some good examples of cross-border cooperation. The Arctic Five (Arctic 5), for instance, is a forum for collaboration among the five universities in northern Finland, Sweden and Norway: The University of Oulu, The University of Lapland, Luleå University of Technology, UiT – The Arctic University of Norway and Umeå University. For more information, see <https://www oulu.fi/thuleinstitute/node/50198>. Another good example of cross-border cooperation is the report *The Cross-Border Cooperation on Innovation – A Joint Taskforce* (University of Oulu, 2019), which not only identifies the main characteristics, the competence areas and the actors in the innovation ecosystems in the Arctic 5 regions, but also makes recommendations for reinforcing both national and international cooperation among the Arctic 5 cities. For further information, see <https://www oulu.fi/oulubusinessschool/node/58610>.
4. The project Cross-Border Cooperation on Innovation – A Joint Taskforce shows that there is ongoing cooperation among the Arctic 5 regions. However, all these projects are either bi- or trilateral.
5. Scopus is the largest abstract and citation database for peer-reviewed literature in the world. It also has the highest limits for one-time downloads of abstract information (2000 abstracts). Research in the social sciences and humanities is underrepresented on Scopus (Martín-Martín et al., 2018) compared with Google Scholar. However, these fields are not the main focus of Smart Specialisation. In Google Scholar and Web of Science, the batch export functionalities are limited, which causes them to become less usable for our purposes. In addition, there are not yet scalable methods to extract data from Google Scholar, nor are there sufficiently nuanced metadata. This makes Google Scholar unsuitable for large-scale, big data abstract text analyses. Research that appears in the grey literature was excluded, as the grey literature is not indexed in the same way as peer-reviewed studies, and selecting and searching for the relevant grey literature introduces additional bias due to human subjectivity in the search and retrieval.
6. Cosine similarity is measured by the cosine of the angle between the probability vectors (which measure the topics) and determines whether two vectors are

pointing in roughly the same direction. The cosine similarity can be represented as  $\frac{A \cdot B}{\|A\| \|B\|}$ , where A and B are the noun probability vectors of a topic for a pair of universities.

7. See note 1.

## ORCID

Moilanen Mikko  <http://orcid.org/0000-0002-7525-860X>

Østbye Stein  <http://orcid.org/0000-0003-3852-7281>

Simonen Jaakko  <http://orcid.org/0000-0001-9528-279X>

## REFERENCES

- Arnold, C. W., Oh, A., Chen, S., & Speier, W. (2015). Evaluating topic model interpretability from a primary care physician perspective. *Computer Methods and Programs in Biomedicine*, 124, 67–75. <https://doi.org/10.1016/j.cmpb.2015.10.014>
- Balland, P.-A., Boschma, R., Crespo, J., & Rigby, D. L. (2019). Smart specialization policy in the European Union: Relatedness, knowledge complexity and regional diversification. *Regional Studies*, 53(9), 1252–1268. <https://doi.org/10.1080/00343404.2018.1437900>
- Bischof, J. M., & Airoidi, E. M. (2012). *Summarizing topical content with word frequency and exclusivity*. In Proceedings of the 29th international conference on machine learning, Edinburgh, UK, June 26–July 1, 2012.
- Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 1(1), 17–35. <https://doi.org/10.1214/07-aos114>
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3, 993–1022.
- Boschma, R. A., & Frenken, K. (2009). Technological relatedness and regional branching. In H. Bathelt, M. P. Feldman, & D. F. Kogler (Eds.), *Beyond territory: Dynamic geographies of innovation and knowledge creation* (pp. 64–81). Routledge.
- Bromme, R. (2000). Beyond one's own perspective: The psychology of cognitive interdisciplinarity. In P. Weingart & N. Stehr (Eds.), *Practicing interdisciplinarity* (pp. 115–133). Toronto University Press.
- Capello, R., & Kroll, H. (2016). From theory to practice in Smart Specialization Strategy: Emerging limits and possible future trajectories. *European Planning Studies*, 24(8), 1393–1406. <https://doi.org/10.1080/09654313.2016.1156058>
- Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (pp. 288–296). MIT Press.
- Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, 17(1), 61–84. <https://doi.org/10.3166/dn.17.1.61-84>
- DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics*, 41(6), 570–606. <https://doi.org/10.1016/j.poetic.2013.08.004>
- European Commission. (2014). *The role of universities and research organisations as drivers for Smart Specialisation at regional level*. [http://ec.europa.eu/research/regions/pdf/publications/Expert\\_Report-Universities\\_and\\_Smart\\_Spec-WebPublication-A4.pdf](http://ec.europa.eu/research/regions/pdf/publications/Expert_Report-Universities_and_Smart_Spec-WebPublication-A4.pdf)

- Foray, D. (2013). The economic fundamentals of smart specialisation. *Ekonomiaz*, 83(2), 83–102. <https://doi.org/10.1016/B978-0-12-804137-6.00002-4>
- Foray, D. (2014). From smart specialisation to smart specialisation policy. *European Journal of Innovation Management*, 17(4), 492–507. <https://doi.org/10.1108/EJIM-09-2014-0096>
- Foray, D., David, P. A., & Hall, B. (2009). Smart specialisation—the concept. *Knowledge Economists Policy Brief*, 9, 100. [https://doi.org/10.1016/S2212-5671\(12\)00146-3](https://doi.org/10.1016/S2212-5671(12)00146-3)
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297. <https://doi.org/10.1093/pan/mps028>
- Huber, F. (2012). On the role and interrelationship of spatial, social and cognitive proximity: Personal knowledge relationships of R&D workers in the Cambridge information technology cluster. *Regional Studies*, 46(9), 1169–1182. <https://doi.org/10.1080/00343404.2011.569539>
- Karlsdóttir, A., & Greve Harbo, L. (2017). *Nordic Arctic strategies in overview*. <https://archive.nordregio.se/Global/Publications/Publications%202017/PB%202017%201.pdf>
- Kristensen, I., Teräs, J., & Wøien, M. (2018). *The potential for smart specialisation for enhancing innovation and resilience in Nordic regions. Preliminary report: Policy and literature review*. <https://www.nordregio.org/wp-content/uploads/2018/02/The-potential-of-Smart-Specialisation-for-enhancing-innovation-and-resilience-in-Nordic-regions-1.pdf>
- Lundquist, K.-J., & Trippel, M. (2013). Distance, proximity and types of cross-border innovation systems: A conceptual analysis. *Regional Studies*, 47(3), 450–460. <https://doi.org/10.1080/00343404.2011.560933>
- Makkonen, T., Weidenfeld, A., & Williams, A. M. (2017). Cross-border regional innovation system integration: An analytical framework. *Tijdschrift voor Economische en Sociale Geografie*, 108(6), 805–820. <https://doi.org/10.1111/tesg.12223>
- Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, web of science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4), 1160–1177. <https://doi.org/10.1016/j.joi.2018.09.002>
- McCann, P., & Ortega-Argilés, R. (2014). Smart specialisation in European regions: Issues of strategy, institutions and implementation. *European Journal of Innovation Management*, 17(4), 409–427. <https://doi.org/10.1108/EJIM-05-2014-0052>
- McCann, P., & Ortega-Argilés, R. (2015). Smart specialization, regional growth and applications to European Union Cohesion Policy. *Regional Studies*, 49(8), 1291–1302. <https://doi.org/10.1080/00343404.2013.799769>
- Muller, E., Zenker, A., Hufnagl, M., Héraud, J.-A., Schnabl, E., Makkonen, T., & Kroll, H. (2017). Smart specialisation strategies and cross-border integration of regional innovation systems: Policy dynamics and challenges for the Upper Rhine. *Environment and Planning C: Politics and Space*, 35(4), 684–702. <https://doi.org/10.1177/0263774X16688472>
- Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 1–15. <https://doi.org/10.1103/physrev.69.026113>
- Oliveira, E. (2015). Constructing regional advantage in branding the cross-border Euroregion Galicia–northern Portugal. *Regional Studies, Regional Science*, 2(1), 341–349. <https://doi.org/10.1080/21681376.2015.1044020>
- Papagiannidis, S., See-To, E. W. K., Assimakopoulos, D. G., & Yang, Y. (2018). Identifying industrial clusters with a novel big-data methodology: Are SIC codes (not) fit for purpose in the internet age? *Computers & Operations Research*, 98, 355–366. <https://doi.org/10.1016/j.cor.2017.06.010>
- Pavone, P., Pagliacci, F., Russo, M., & Giorgi, A. (2019). *R&I smart specialisation strategies: classification of EU region 'priorities'. Results from automatic text analysis*. <https://iris.unimore.it/retrieve/handle/11380/1196211/252292/0148.pdf>
- Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515), 988–1003. <https://doi.org/10.1080/01621459.2016.1141684>
- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). *Stm: An R package for structural topic models*. *Journal of Statistical Software*, 91(2), 40. <https://doi.org/10.18637/jss.v091.i02>
- University of Oulu. (2019). *The cross-border cooperation on innovation – A joint taskforce*. <https://www oulu.fi/oulubusinessschool/node/58610>
- van Ark, B., O'Mahony, M., & Timmer, M. P. (2008). The productivity gap between Europe and the United States: Trends and causes. *Journal of Economic Perspectives*, 22(1), 25–44. <https://doi.org/10.1257/jep.22.1.25>