@rural_gis

Exploring the thematic evolution of geographical information science and systems research with topic modelling

Robert Berry[1] , Scott Orford[2], Lucy Clarke[3], & Caitlin Hafferty[4]

[1]Countryside and Community Research Institute (CCRI), University of Gloucestershire
[2]School of Geography and Planning, Cardiff University
[3]School of Education and Applied Sciences, University of Gloucestershire
[4]Environmental Change Institute, Oxford University

G.I.S

Computer systems for managing, analysing and visualising geographic data

Geographical Information **Science** and **Systems**

Science (GISci) - fundamental theories, principles, methods, and knowledge

Systems (GIS) - technology of problem solving (tools and their applications)

Longley et al., 2015

Evolution of GIS research (GISR)

- Large gaps in our understanding
- Focus on origins and tech dev
- Less so **thematic** development
- Bibliographic data analyses of GISR literature limited

"...few comprehensive sources that cover the full range of GIS [research] (thematically, spatially, and temporally)..."

Chrisman (2016, p.1)

Bibliographic analyses of GISR

- Predominantly scientometric
- Limited thematic analysis
- Focussed on specific area of lit
- Lit searches limited scope (small samples)
- Lack of data filtering/cleaning
- No differentiation of GIS and GISci

Author(s)	Time range	n (articles)
Wu et al (2023)	1991-2020	9,400 (10 journals)
Huang (2022)	1990-2017	16,096
de Melo and Queiroz (2019)	2007-2016	2,053
Biljecki (2016)	2000-2014	12,436
Duckham (2015)	< 2015	27 journals
Tian et al (2008)	1997-2006	9,849
Sun and Manson (2011)	1992-2007	20,181

Selected recent articles on bibliographic analysis of GIS literature

Why improving understanding of GISR thematic structure?

1. Better understand how **knowledge has developed over time** in GISR – detecting trends/shifts/“turning points” that have shaped research (Chen, 2004)
2. Help identify **gaps in the literature** and guide research direction- e.g., focussing on unexplored areas (José De Oliveira et al., 2019)
3. Inform decisions regarding **resource allocation** and identifying areas within a field of research that require further investment (De Bellis, 2009)
4. Understand if/how GISR responds to **societal challenges/real-world problems**

Aim

Conduct an extensive, robust computational analysis of the GISR literature to produce a comprehensive model of its thematic structure

Research questions

RQ1 – What is the thematic structure of the GISR literature?

RQ2 – Can GISR lit be meta-categorised into ***science*** and ***systems****?*

RQ2 - How has the thematic structure of GISR evolved over time?

RQ4 – What are the drivers for major changes/turning points in GISR?

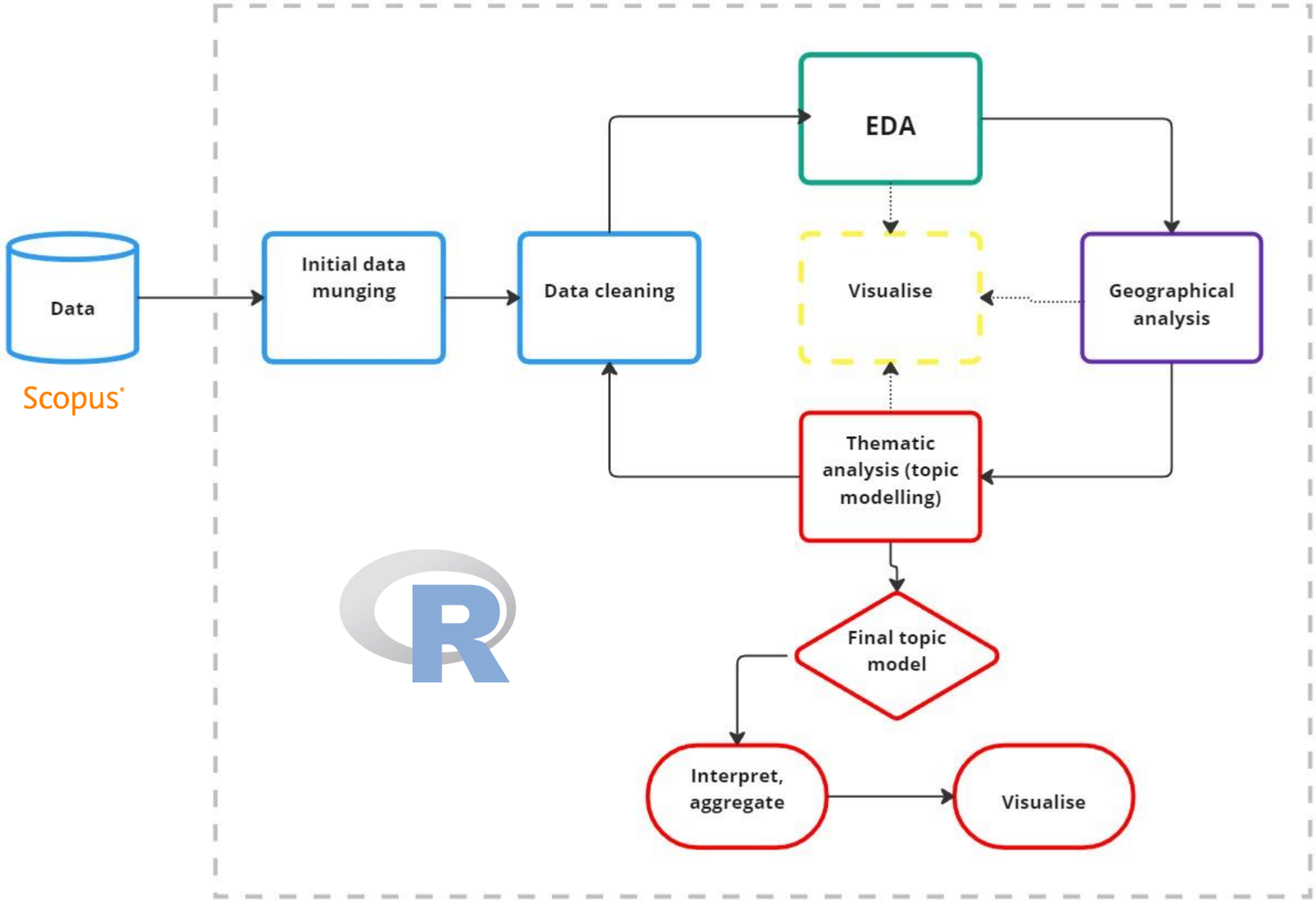
RQ5 – How is GISR responding to major societal challenges?

Methodology

Reproducible research

R programming language (v. 4.2.0)

R Studio (v. 2023.06.0)



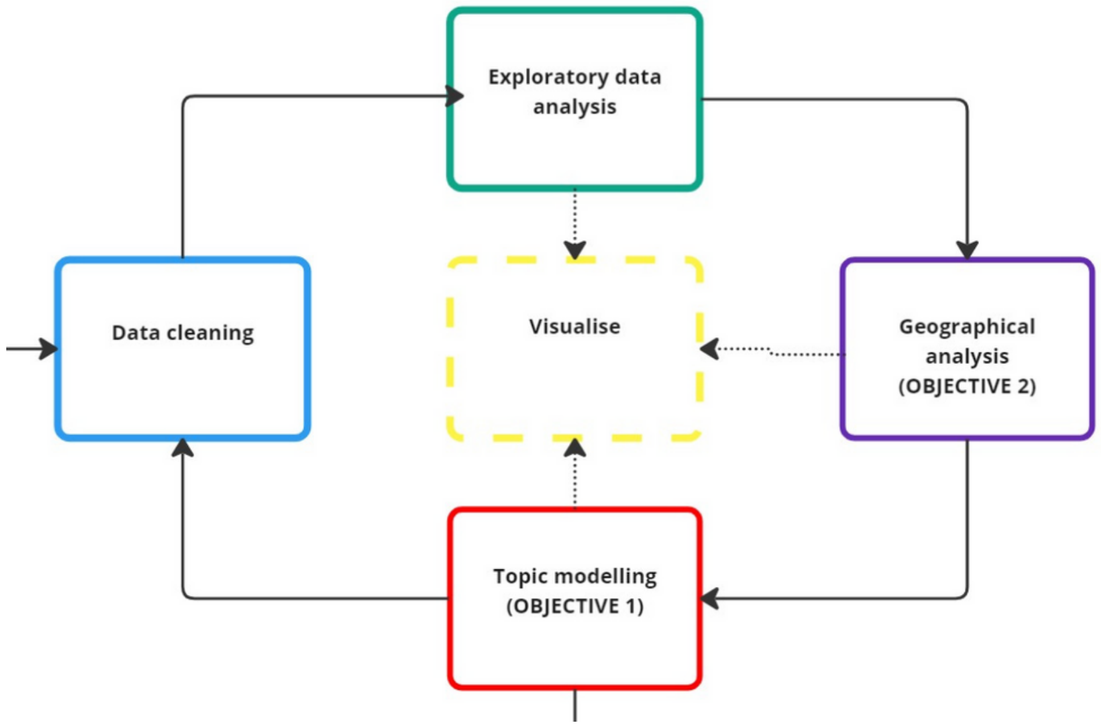
Data acquisition

```
( TITLE-ABS-KEY ( {geographic* information} ) OR
TITLE-ABS-KEY ( {geoinformation} ) OR
TITLE-ABS-KEY ( {geo-information} ) OR
TITLE-ABS-KEY ( {geo-spatial} ) OR
TITLE-ABS-KEY ( {geospatial} ) OR
TITLE-ABS-KEY ( {spatial data} ) OR
TITLE-ABS-KEY ( {geodata} ) OR
TITLE-ABS-KEY ( {geo-data} ) OR
TITLE-ABS-KEY ( {giscience} ) OR
TITLE-ABS-KEY ( {GIS} ) OR
TITLE-ABS-KEY ( {G.I.S} ) OR
TITLE-ABS-KEY ( {spatial analysis} ) OR
TITLE-ABS-KEY ( {gi-science} ) AND
LANGUAGE ( english ) ) AND
PUBYEAR > 1969 AND PUBYEAR < 2022 OR
TITLE-ABS-KEY ({geographic* data}) OR
SRCTITLE(International Journal of Geographical Information Science) OR
SRCTITLE(Transactions in GIS) OR
SRCTITLE(Journal of Spatial Information Science) OR
SRCTITLE(GeoInformatica) AND
( LIMIT-TO ( SRCTYPE,"j" ) ) AND
( LIMIT-TO ( PUBSTAGE,"final" ) ) AND
( LIMIT-TO ( DOCTYPE,"ar" ) ) AND
( LIMIT-TO ( LANGUAGE,"English" ) )
```

ELSEVIER Scopus®

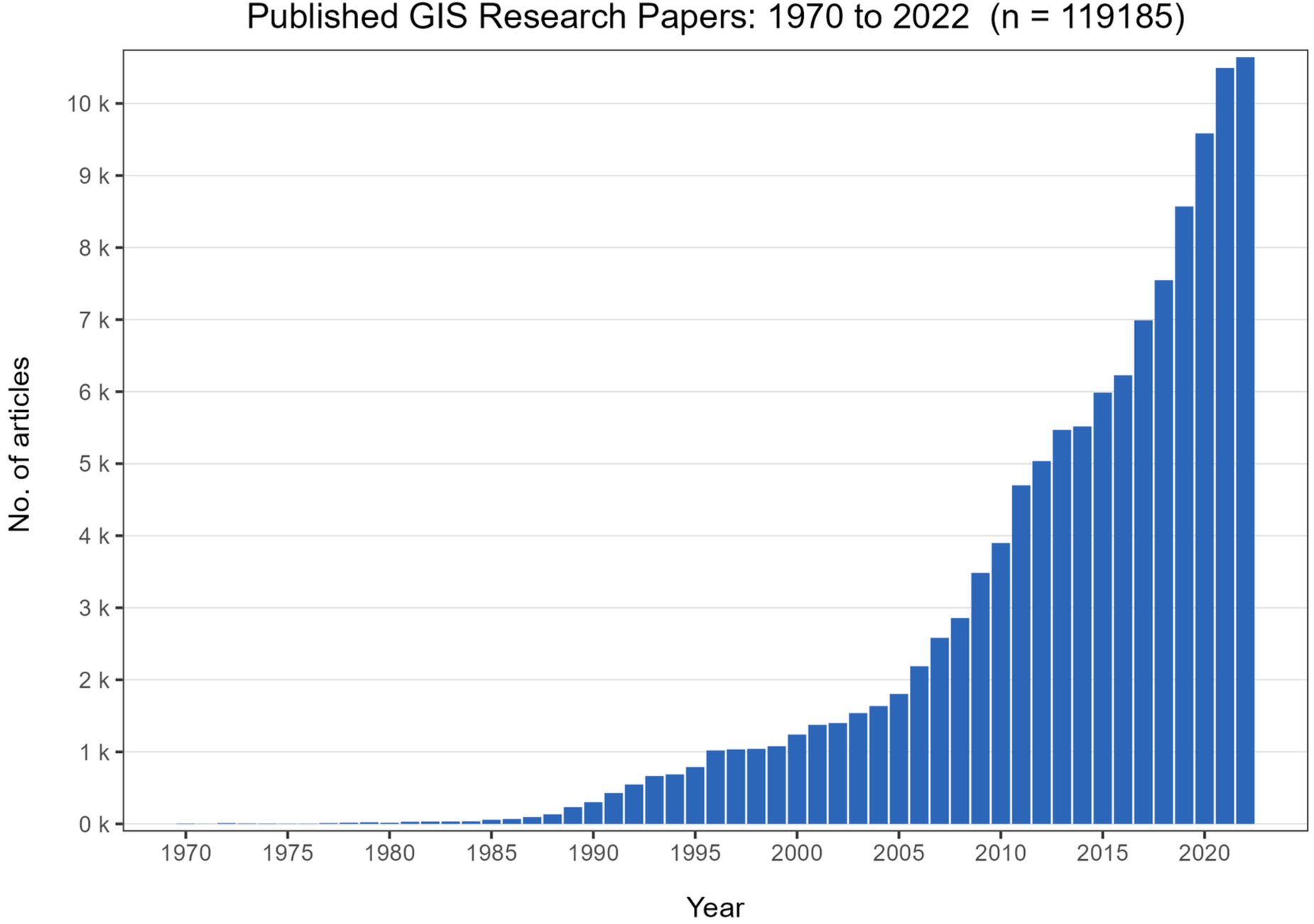
- “Topic” search
  - *Title*, *Keywords*, *Abstract*
- Includes core GIS journals (Duckham, 2016)
- **139,491** articles from **8,354** journals

Data cleaning



- Extensive, iterative
- Removed:
  - 1,077 journals
  - 20,367 papers
- Cleaned DB:
  - **119,185 papers**
  - **>8000 journals**

EDA

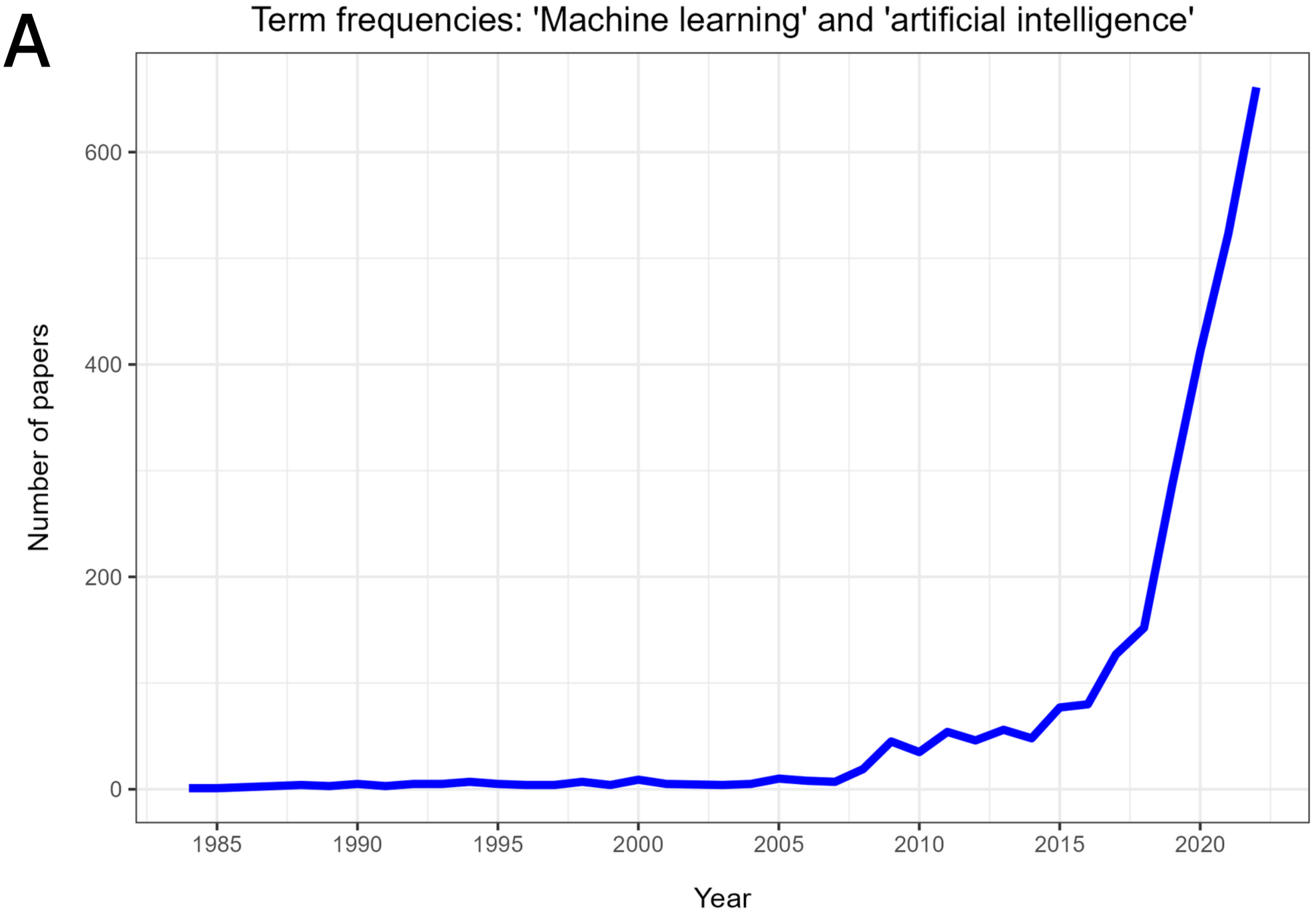

Published GIS Research Papers: 1970 to 2022 (n = 119185)
No. of articles
10 k
9 k
8 k
7 k
6 k
5 k
4 k
3 k
2 k
1 k
0 k
1970
1975
1980
1985
1990
1995
2000
2005
2010
2015
2020
Year

EDA

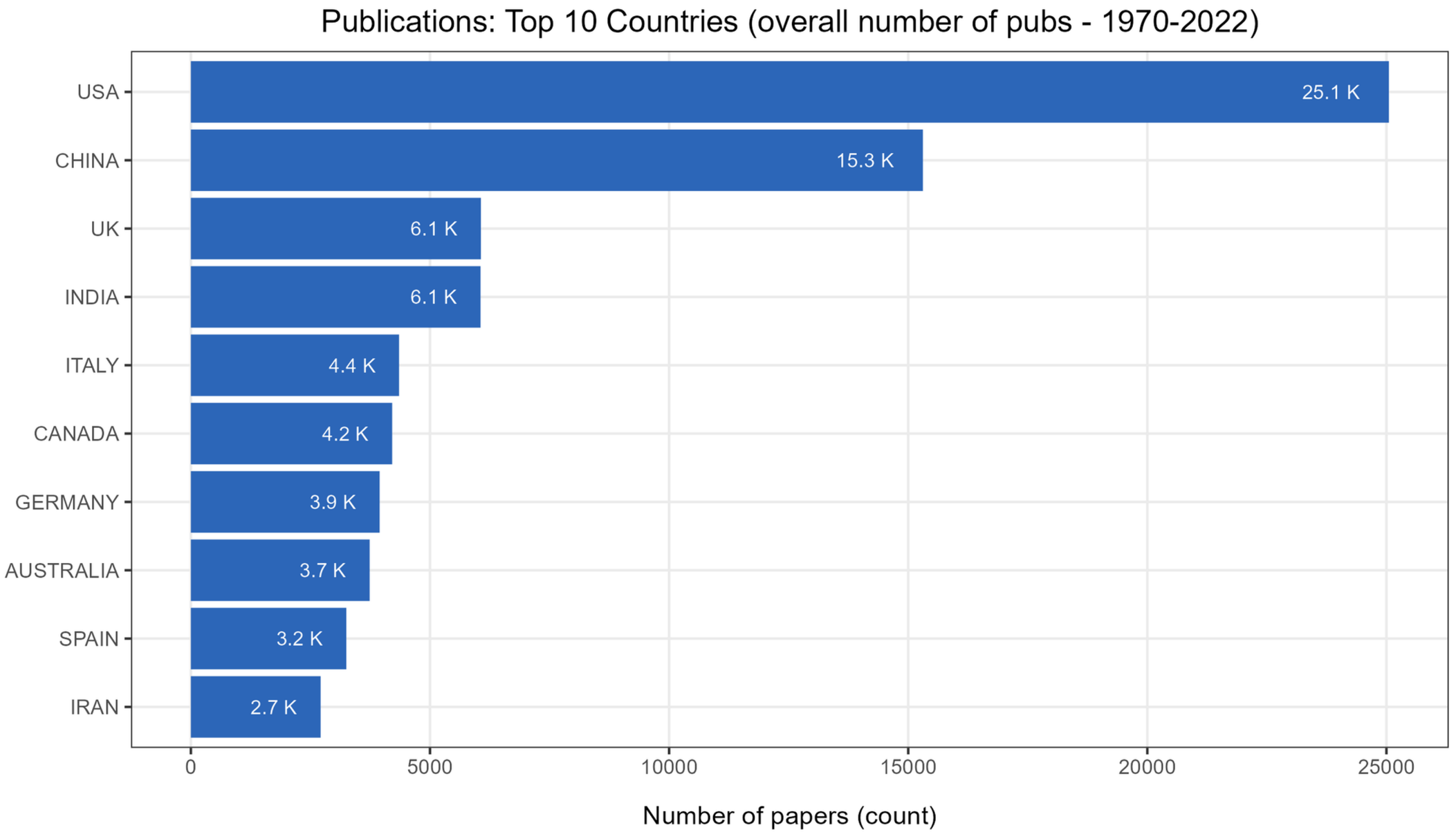
Title	Author(s)	Year	Journal	Citations
Very high-resolution interpolated climate surfaces for global land areas	Hijmans et al	2005	International Journal of Climatology	**14715**
Local indicators of spatial association - LISA	Anselin	1995	Geographical Analysis	**6580**
The global distribution and burden of dengue	Bhatt et al	2013	Nature	**5516**
Koppen's climate classification map for Brazil	Alvares et al	2013	Meteorologische Zeitschrift	**5447**
Large area hydrologic modeling and assessment part i: model development	Arnold et al	1998	Journal of the American Water Resources Association	**5284**
A review of assessing the accuracy of classifications of remotely sensed data	Elly	1991	Remote Sensing of Environment	**5261**
Predictive habitat distribution models in ecology	Guisan & Zimmermann	2000	Ecological Modelling	**5210**
Collinearity: a review of methods to deal with it and a simulation study evaluating their performance	Dormann et al	2013	Ecography	**4669**
The shuttle radar topography mission (SRTM)	Farr et al	2007	Reviews of Geophysics	**4415**
A global map of human impact on marine ecosystems	Halpern et al	2008	Science	**4268**

Top ten articles (in the geographical information science & systems literature) by number of citations (Data source: Scopus, January 2023)

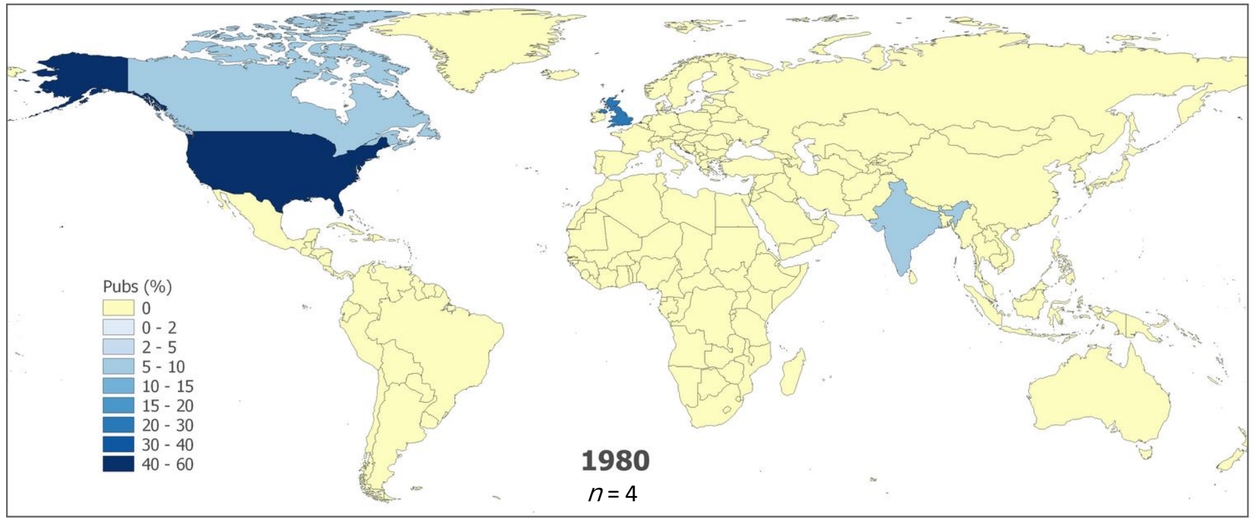
EDA

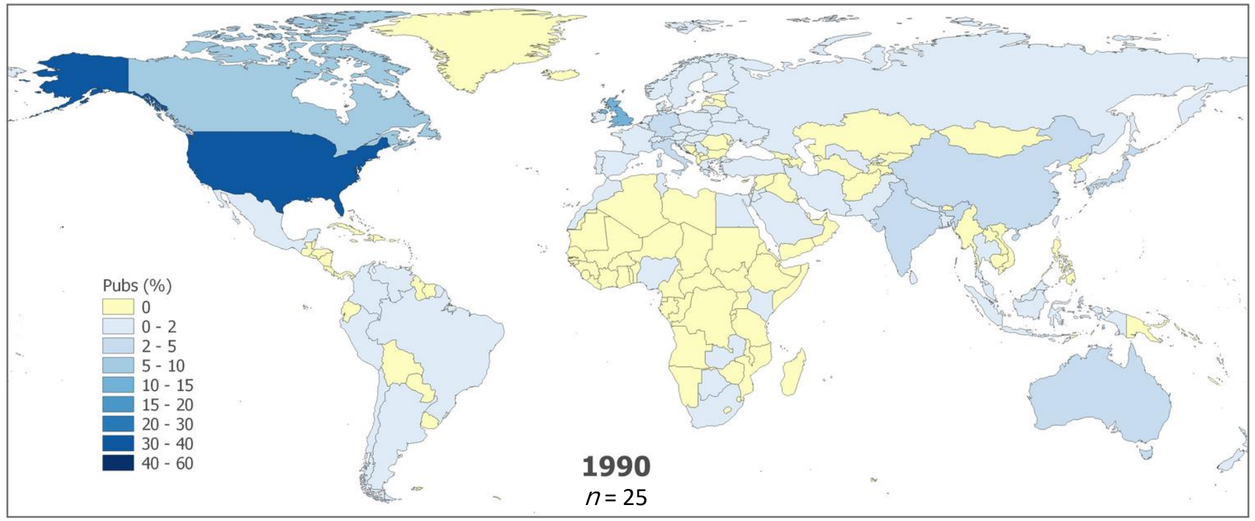

Term frequencies: 'Machine learning' and 'artificial intelligence'
Number of papers
0
200
400
600
1985
1990
1995
2000
2005
2010
2015
2020
Year

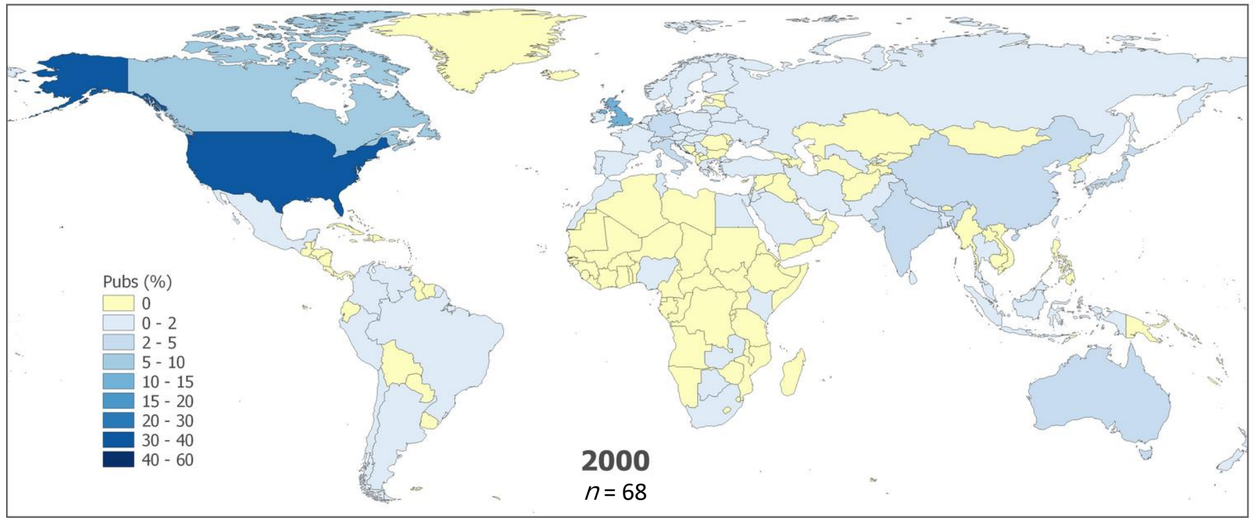
Geographical analysis

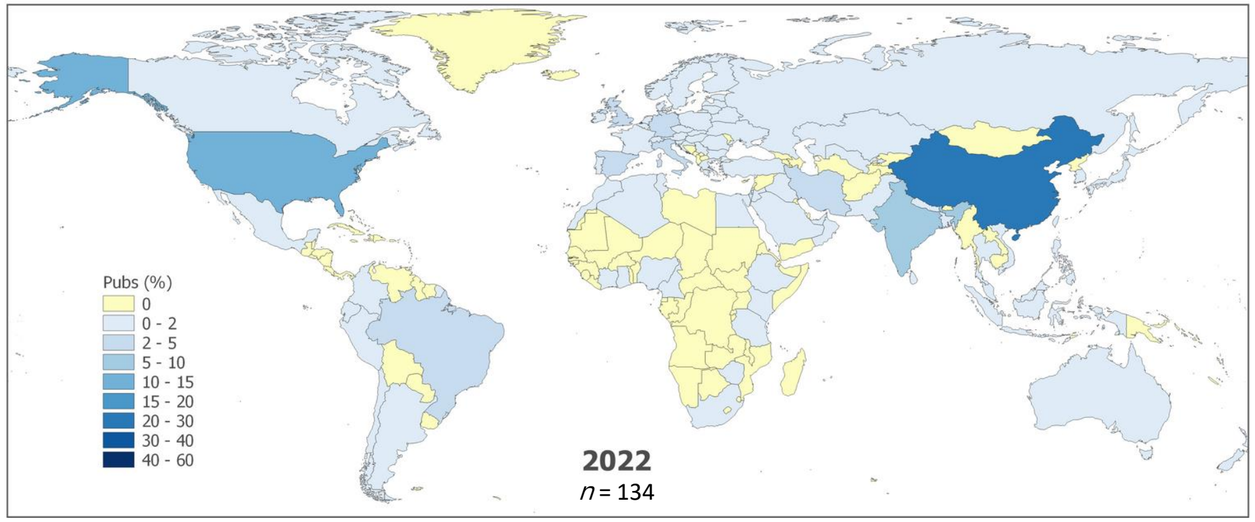


Geographical analysis

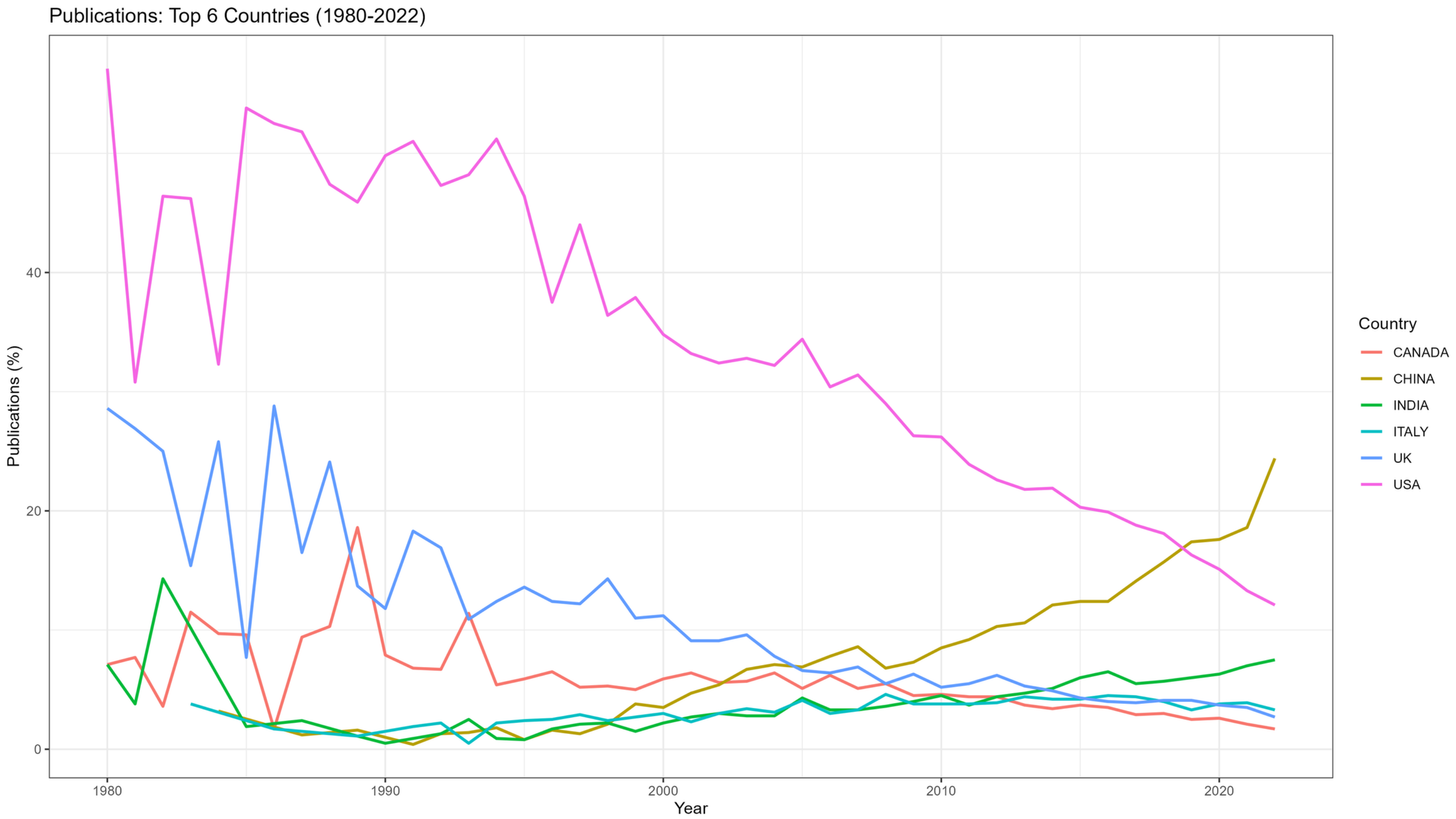








Geographical analysis

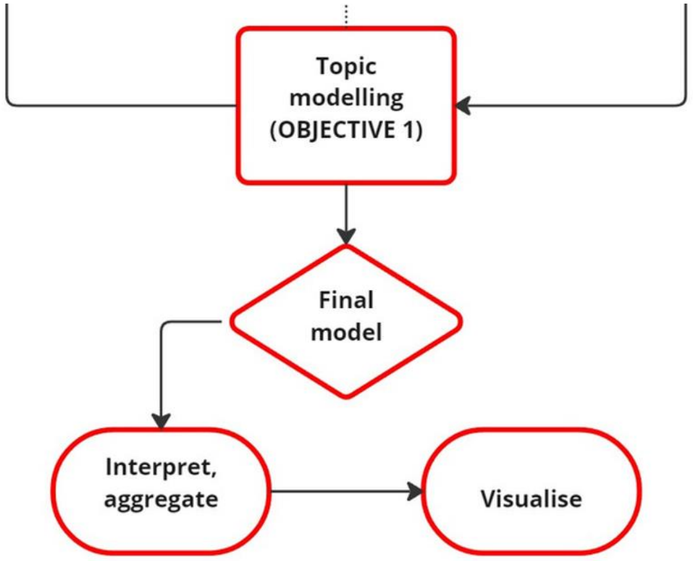


Thematic Analysis

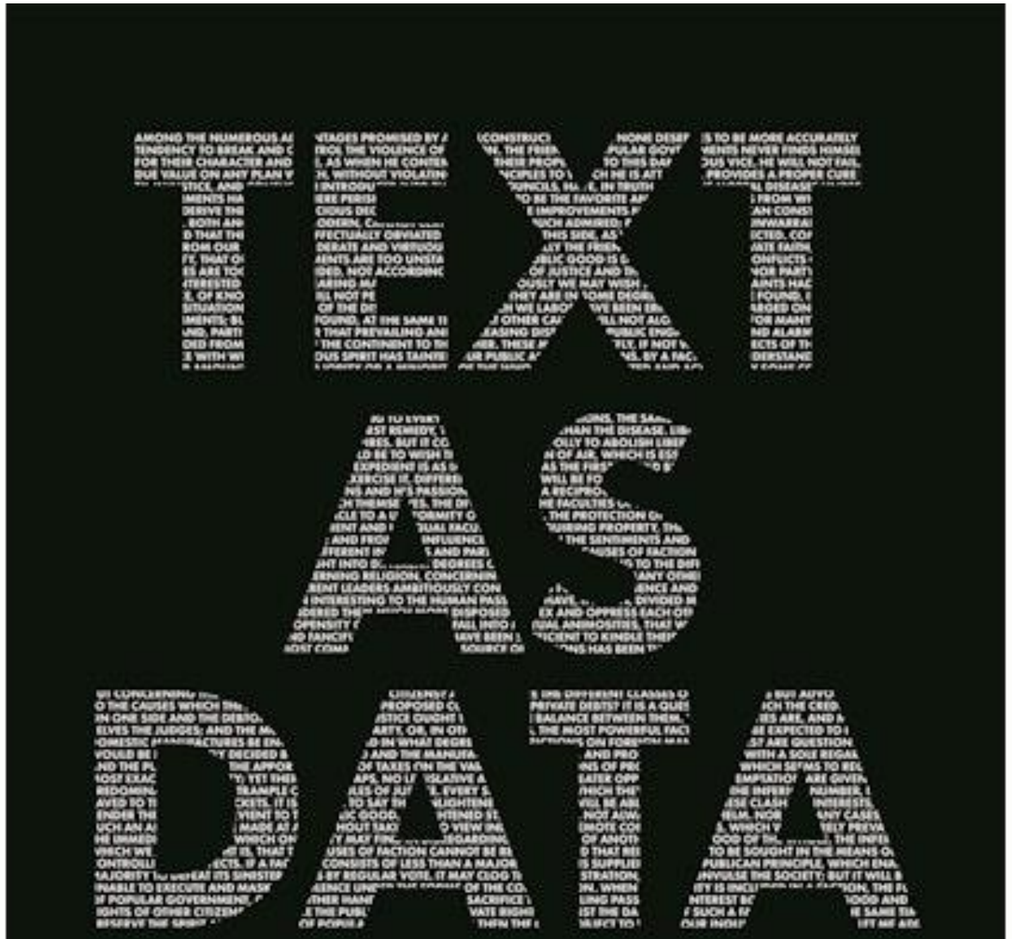
1. Develop a **thematic structure** for GIS lit
2. Chart prevalence of **topics over time**

- Using AI/machine learning (ML)
- Data source → **abstract text**
  - Richer than keywords, more informative than “ready-made” subject area tags

Topic modelling (TM)



- Natural language processing (NLP)
- Computational content analysis
- Unsupervised probabilistic ML
- Discovers latent or “hidden” topics
- For large-huge databases

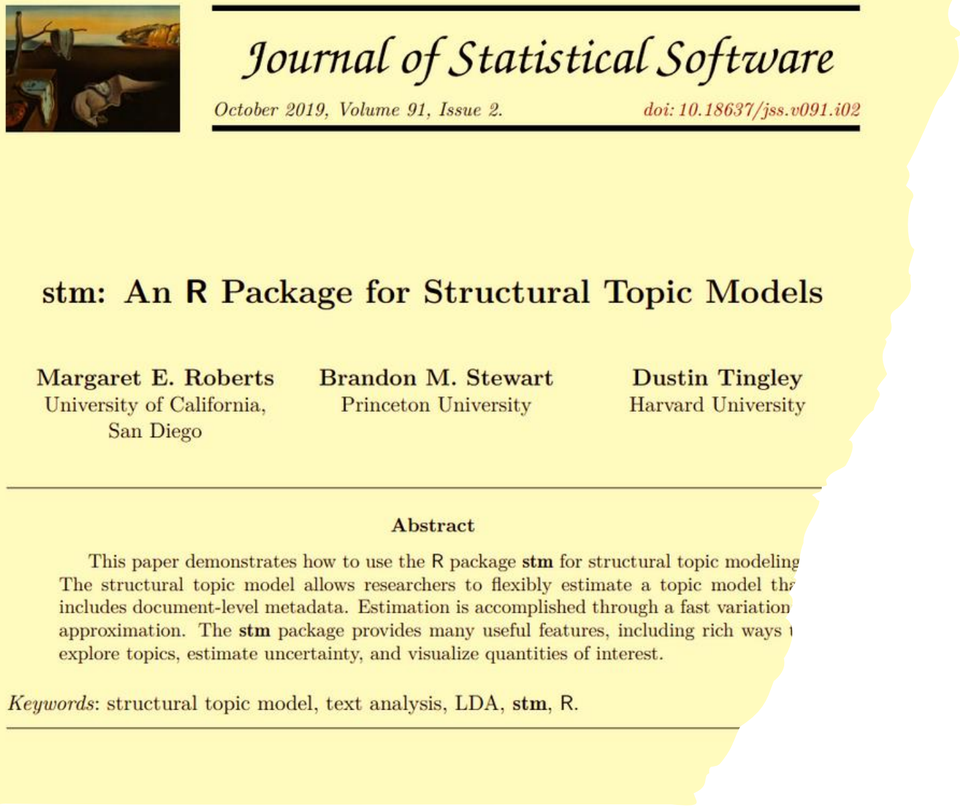


A New Framework for Machine Learning and the Social Sciences

Justin Grimmer | Margaret E. Roberts | Brandon M. Stewart

Structural topic model (STM)

Roberts et al (2019)



Journal of Statistical Software

October 2019, Volume 91, Issue 2. *doi: 10.18637/jss.v091.i02*

stm: An R Package for Structural Topic Models

Margaret E. Roberts
University of California, San Diego

Brandon M. Stewart
Princeton University

Dustin Tingley
Harvard University

Abstract

This paper demonstrates how to use the R package **stm** for structural topic modeling. The structural topic model allows researchers to flexibly estimate a topic model that includes document-level metadata. Estimation is accomplished through a fast variational approximation. The **stm** package provides many useful features, including rich ways to explore topics, estimate uncertainty, and visualize quantities of interest.

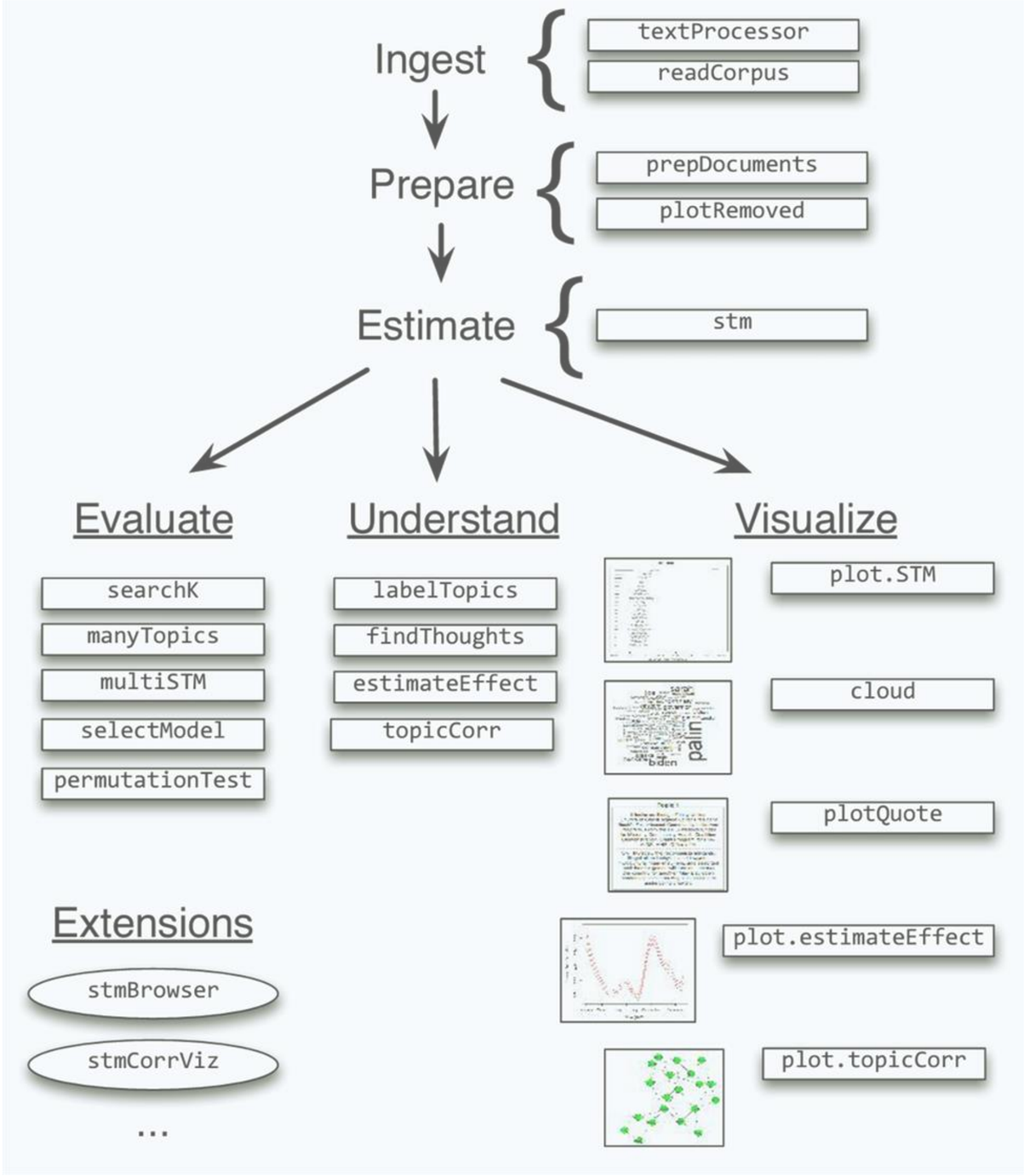
Keywords: structural topic model, text analysis, LDA, **stm**, R.

- State of the art framework for TM
- Estimate effect of metadata on topics
- Diverse applications
  - Political science (Johnston and Sprong, 2023)
  - Human-nature relations (Mul et al., 2022)
  - Land Use Classification (Shao et al., 2020)

structuraltopicmodel.com

The “STM” R package: Functions

(Roberts et al. (2019, p.6)



https://cran.r-project.org/web/packages/stm/index.html

STM: Preparation

- Remove papers where *n* words in abstract < 100
- ***stm::textProcessor***
  - Convert to lower case
  - Remove punctuation
  - Remove stop words (e.g., `and, the, that`)
  - Stem words (e.g., `family, families, family's` mapped to `famili`)
- ***stm::prepDocuments***
  - Several corpus manipulations including removing very rare words

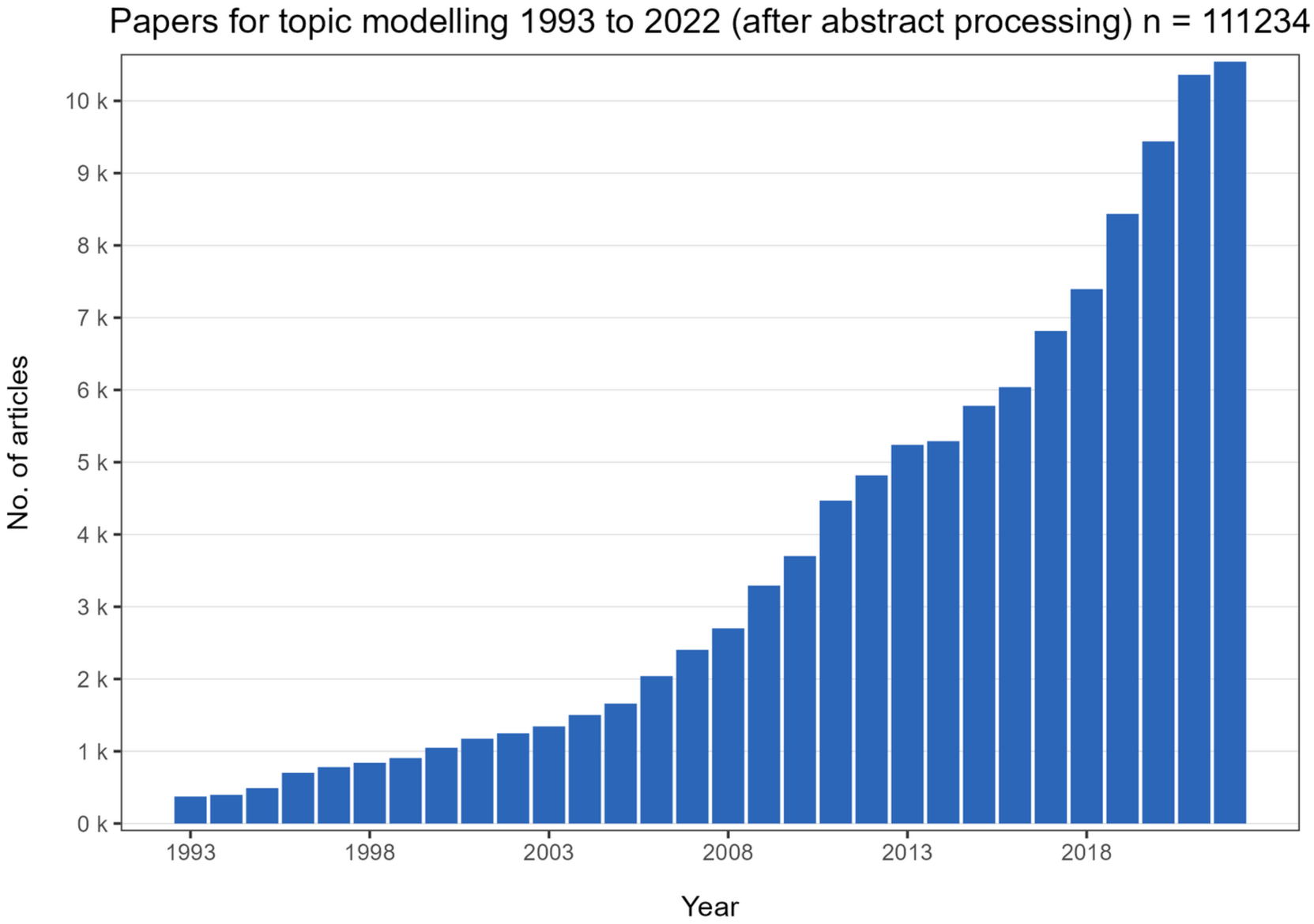
STM: Preparation

Conversion of text into numerical data

- **Document 1**: *"I love cats"*
- **Document 2**: *"I hate dogs"*
- **Document 3**: *"I have a pet cat and a pet dog"*

Document-term matrix

	I	love	cats	hate	dogs	have	a	pet	and
D1	1	1	1	0	0	0	0	0	0
D2	1	0	0	1	1	0	0	0	0
D3	1	0	1	0	0	1	2	2	1

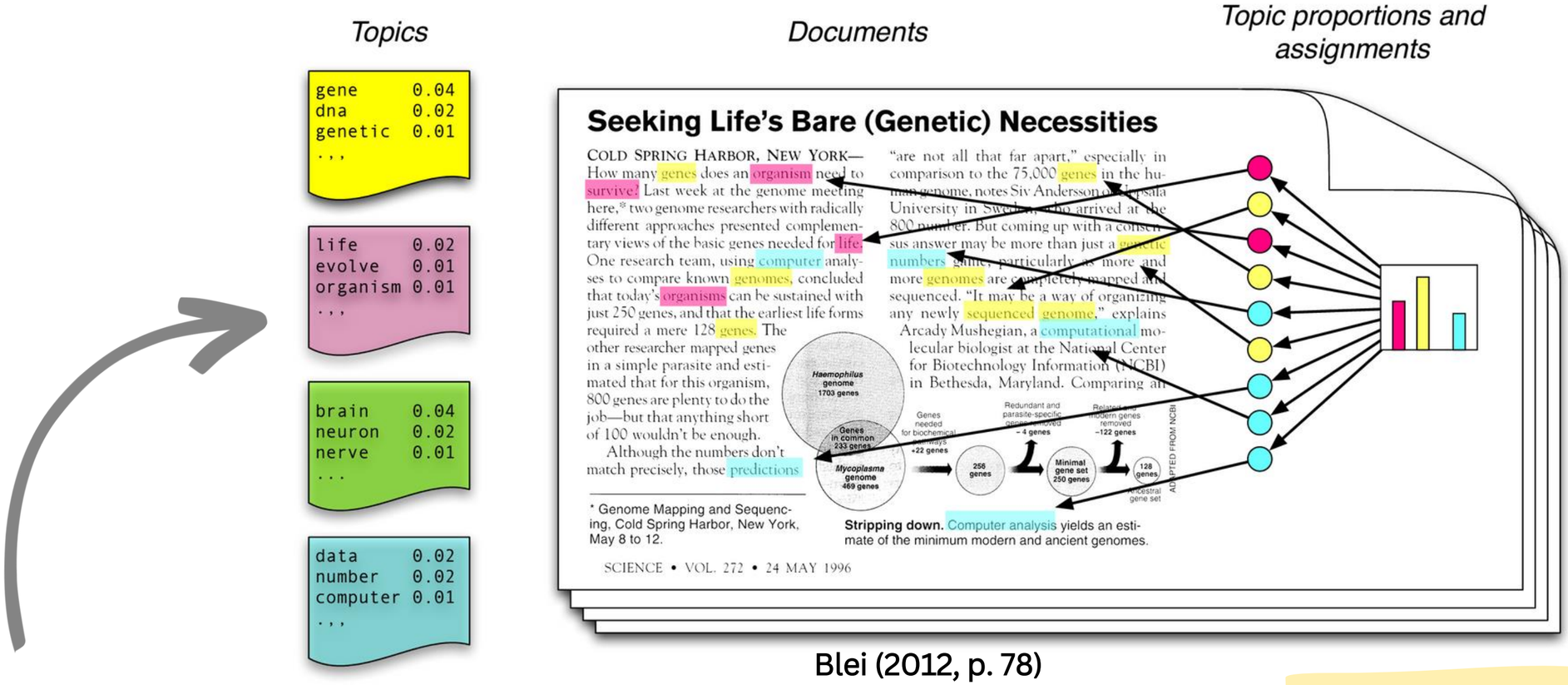


Papers
111K

Corpus
1.9M words

Vocabulary
17K words

Probabilistic topic modelling



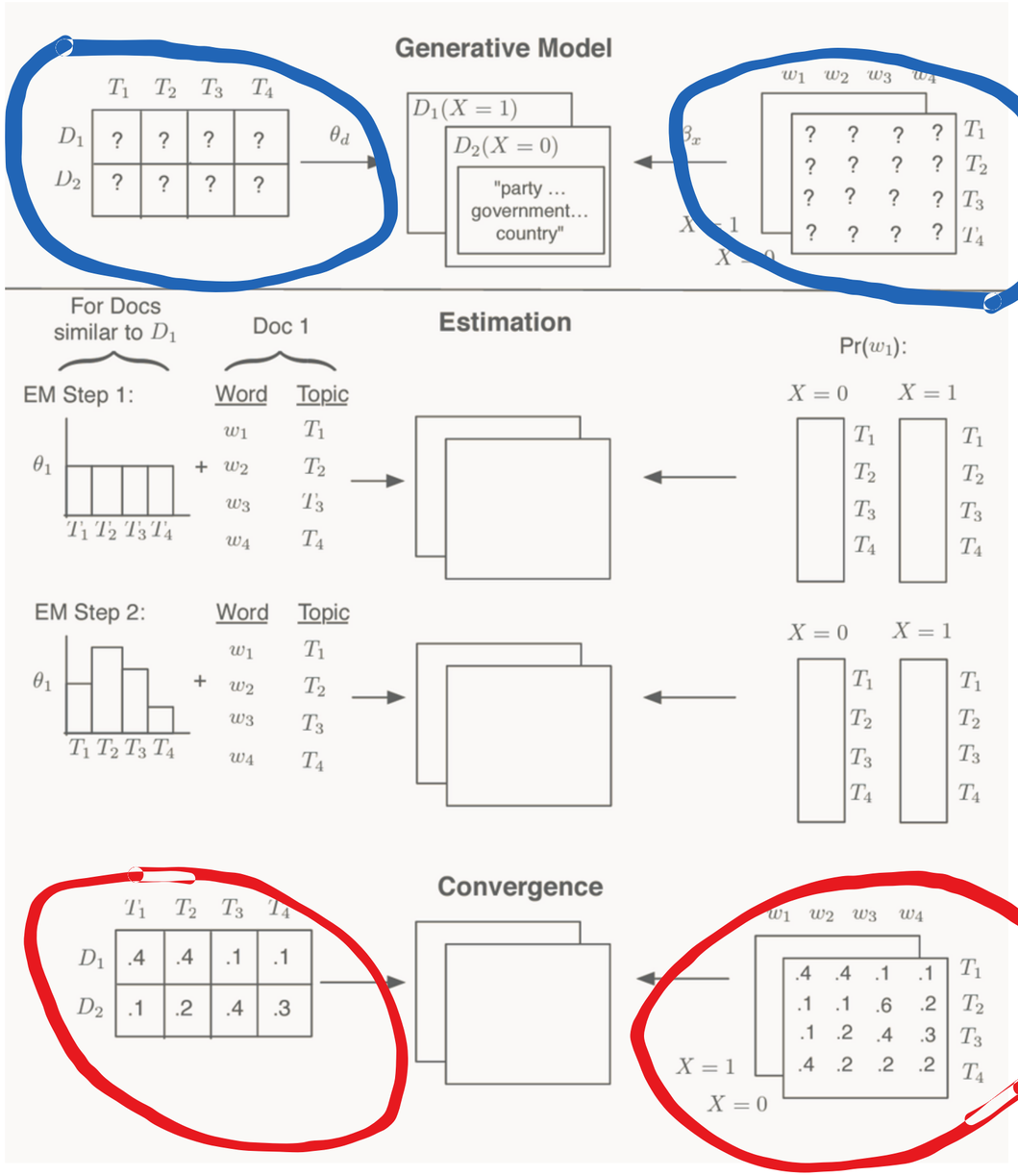
Blei (2012, p. 78)

A **topic** is a probability distribution over **words**

A **document** is a probability distribution over **topics**

Generative process and estimation of the STM

(Roberts et al. (2019, p.4)



Prior distribution

$$\beta_{i,k,v} = \frac{exp\left(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_i,v} + \kappa^{(i)}_{y_i,k,v}\right)}{\sum_v exp\left(m_v + \kappa^{(t)}_{k,v} + \kappa^{(c)}_{y_i,v} + \kappa^{(i)}_{y_i,k,v}\right)}$$

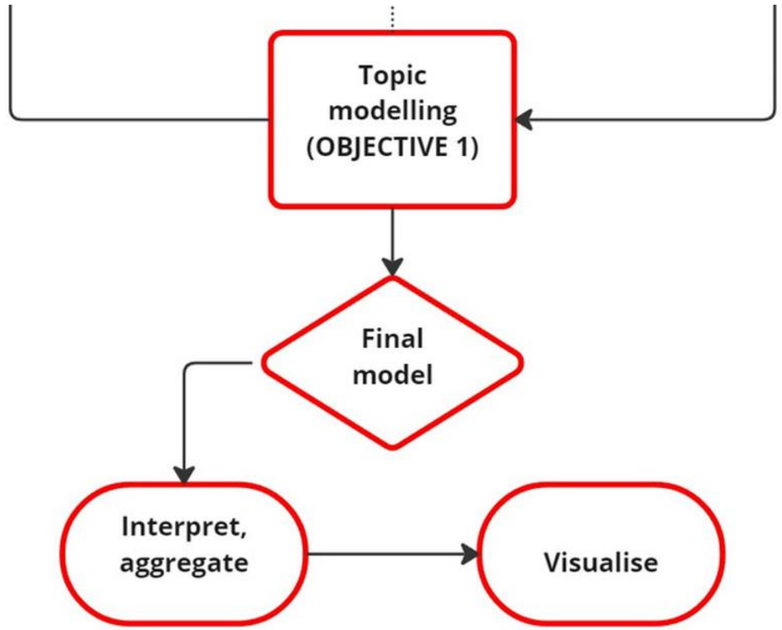
$$P(\eta, z, \kappa', \gamma, \Sigma | W, X, Y, k, m) \propto$$
$$[\prod_{i=1}^{D} Normal(\eta_i | X_i, \gamma, \Sigma)(\prod_{n=1}^{N} Multinomial(z_{n,i} | \Theta_i) x Multinomial(w_n | \beta_{i,k=z_{d,n}}))] x \prod p(\kappa) \prod p(\tau)$$

Posterior distribution

Topic modelling : Challenges

1. Cleaning data
2. Computation
3. Finding *K* (no. of topics)
4. Interpreting topics
5. Aggregating topics

Linked

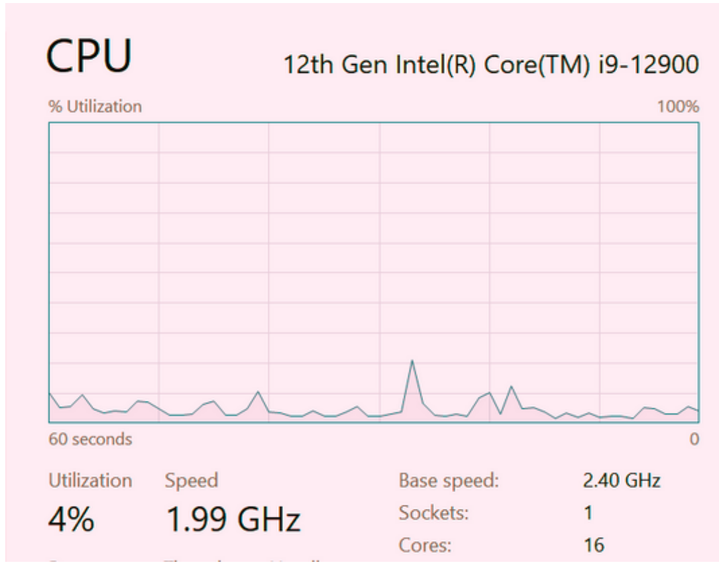


Challenge 1: Cleaning data

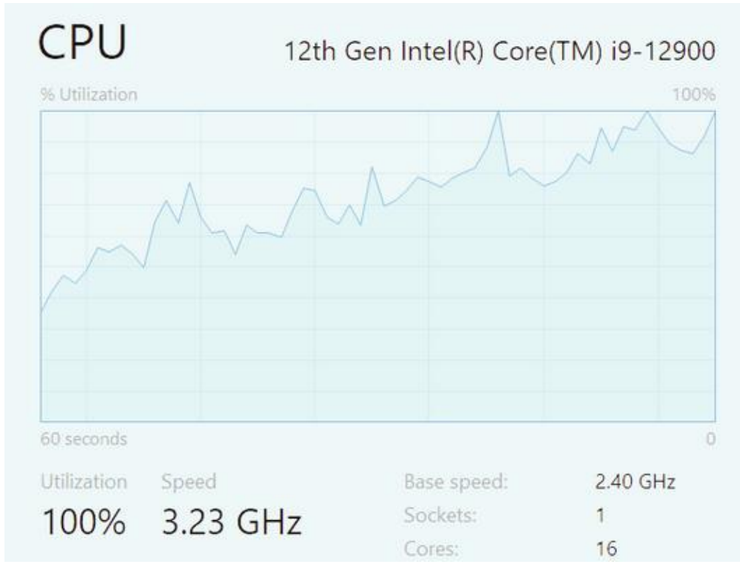
- **2,189 terms** removed through iterative model runs
  - Geographical (continents, countries, regions, languages, nationalities)
  - Common subject terms (e.g., “geographical information systems”)
  - Other stop words not removed in first step
  - Numbers (incl. dates)

Challenge 2: Computation

- Full model estimation run time
  - **35+ hrs** (standard)
  - **4.5 hrs** (parallel)



purrr (standard)



furrr (parallel)

Challenge 3: Finding *K*

How to determine number of topics *K*?

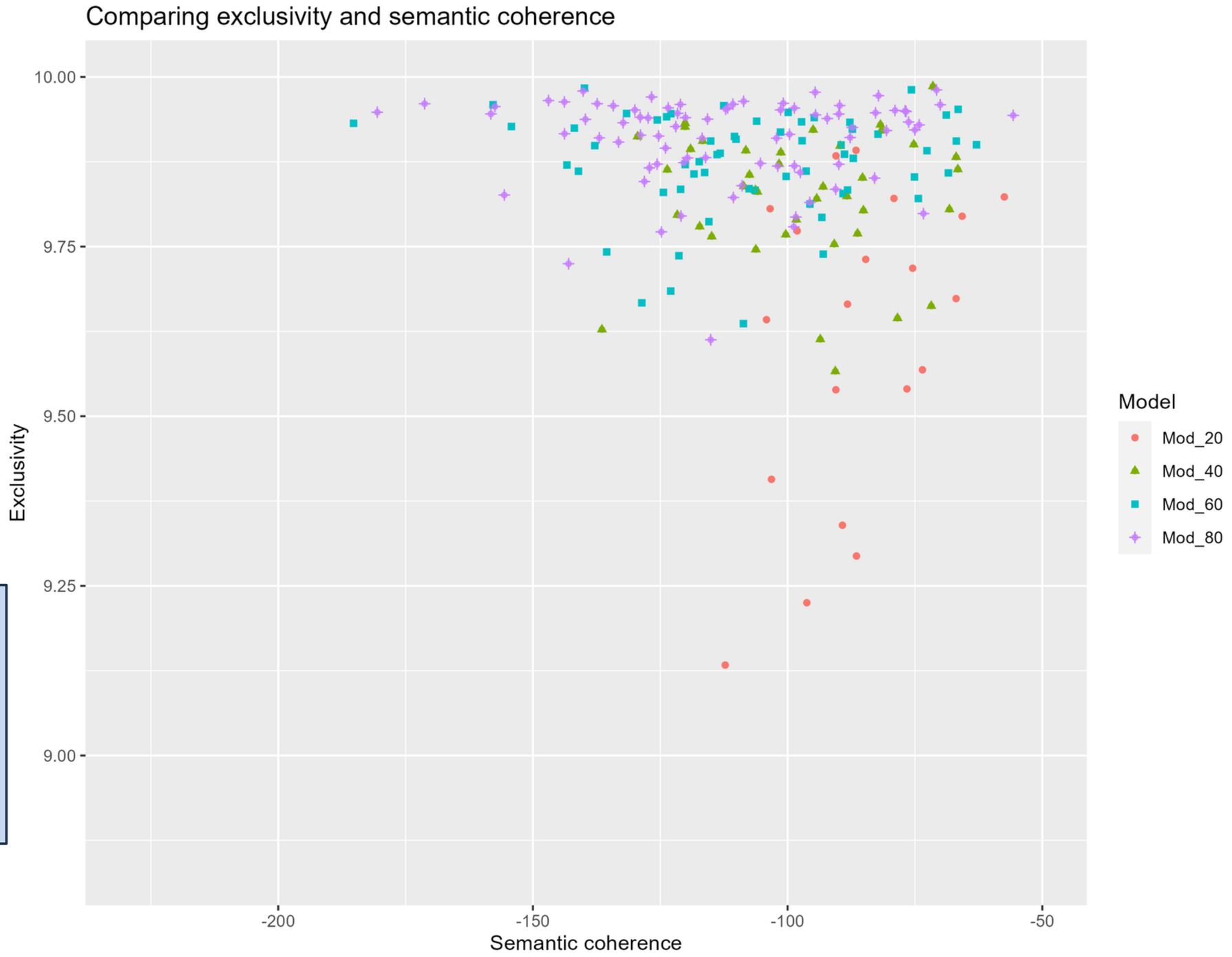
- No “best” model – no “right” answer for *K*
- No systematic evaluation framework
- One solution - compute **multiple models with different K** and evaluate outputs
- Combination of:
  - Quantitative evaluation metrics
  - Qualitative assessment of model outputs
  - Domain knowledge

STM: Finding *K*

Quantitative evaluation

Lee and Minmo (2014)

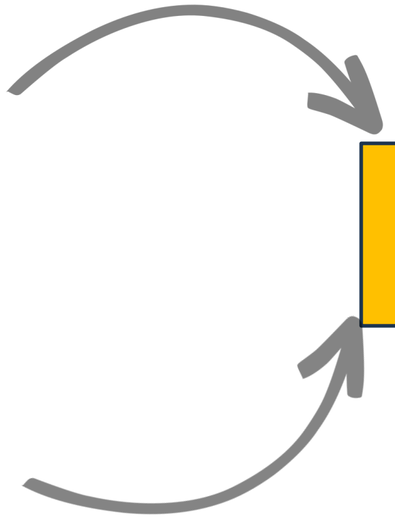
K = 95



Challenges 3 & 4: Finding *K* and interpreting topics

Finding ***K*** and topic interpretation is performed concurrently using range of **quant and qual** STM outputs:

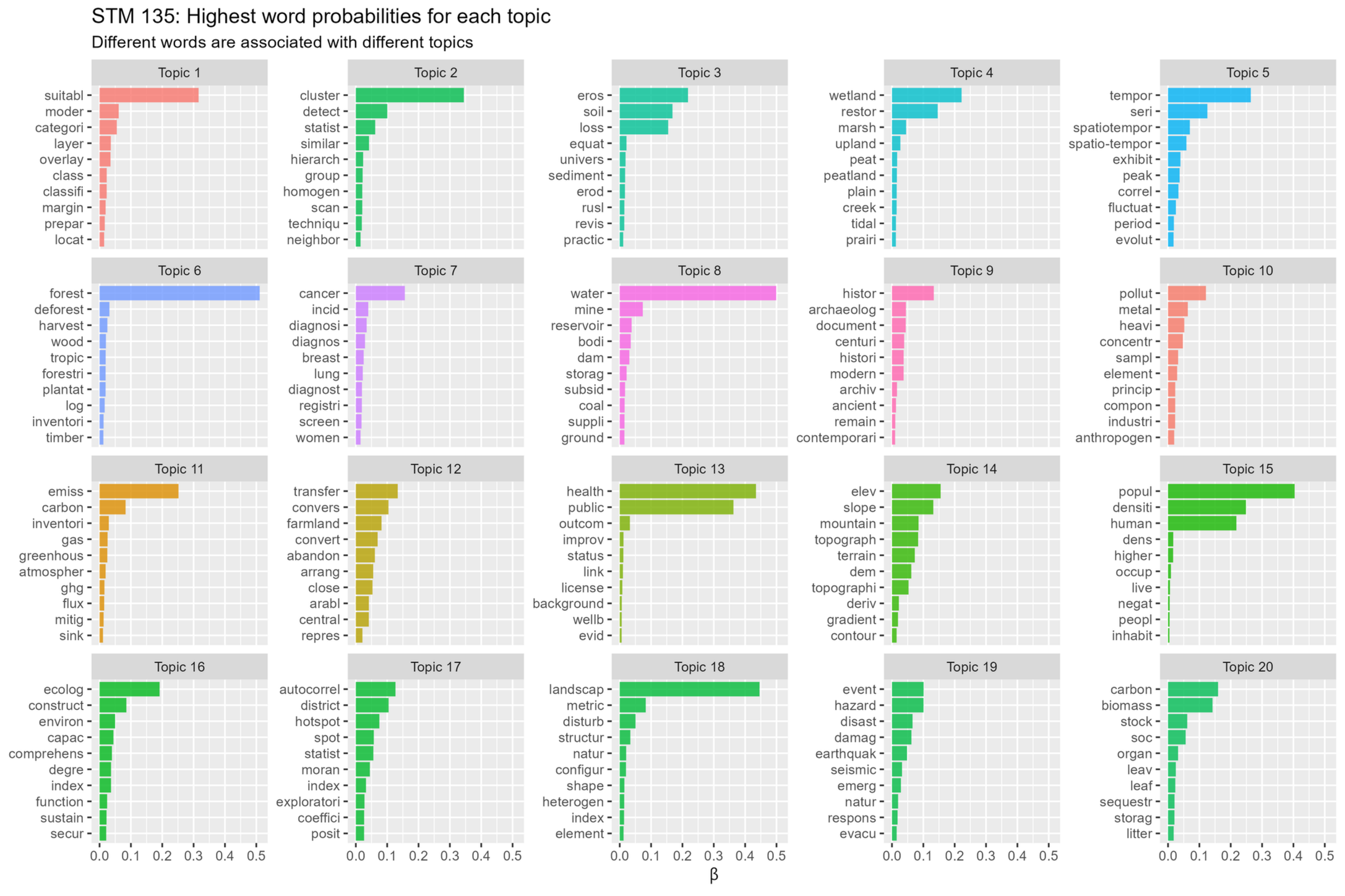
- *labelTopics*
- *Topic plots*
- *ldaViz*
- *wordCloud*
- *findThoughts*
- *Literature sources*



For series of **15 models** (***K*** = 10 to ***K*** = 150)

Final model selected

$K=135$

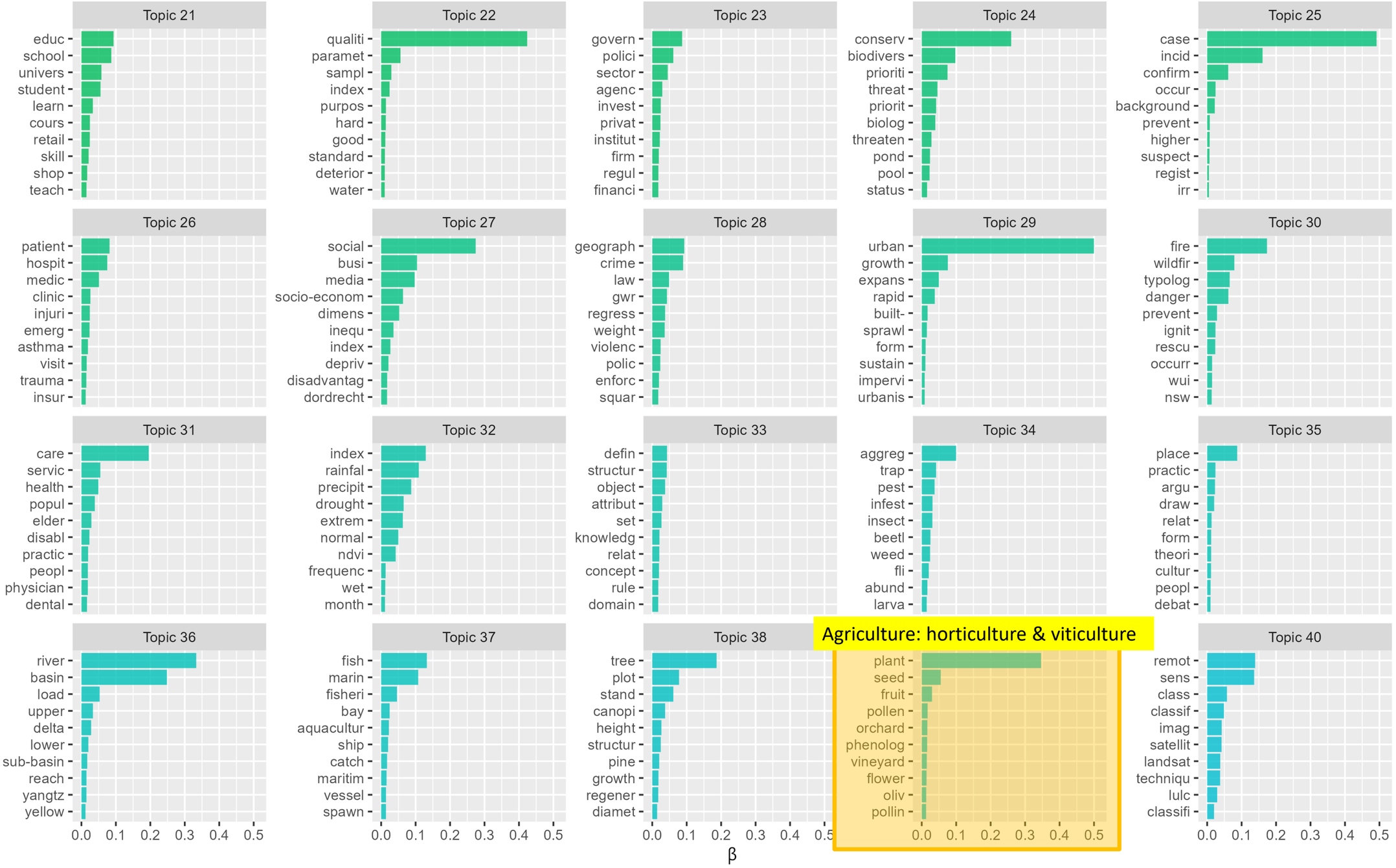


stm::labelTopics

Qualitative topic evaluation

STM 135: Highest word probabilities for each topic

Different words are associated with different topics

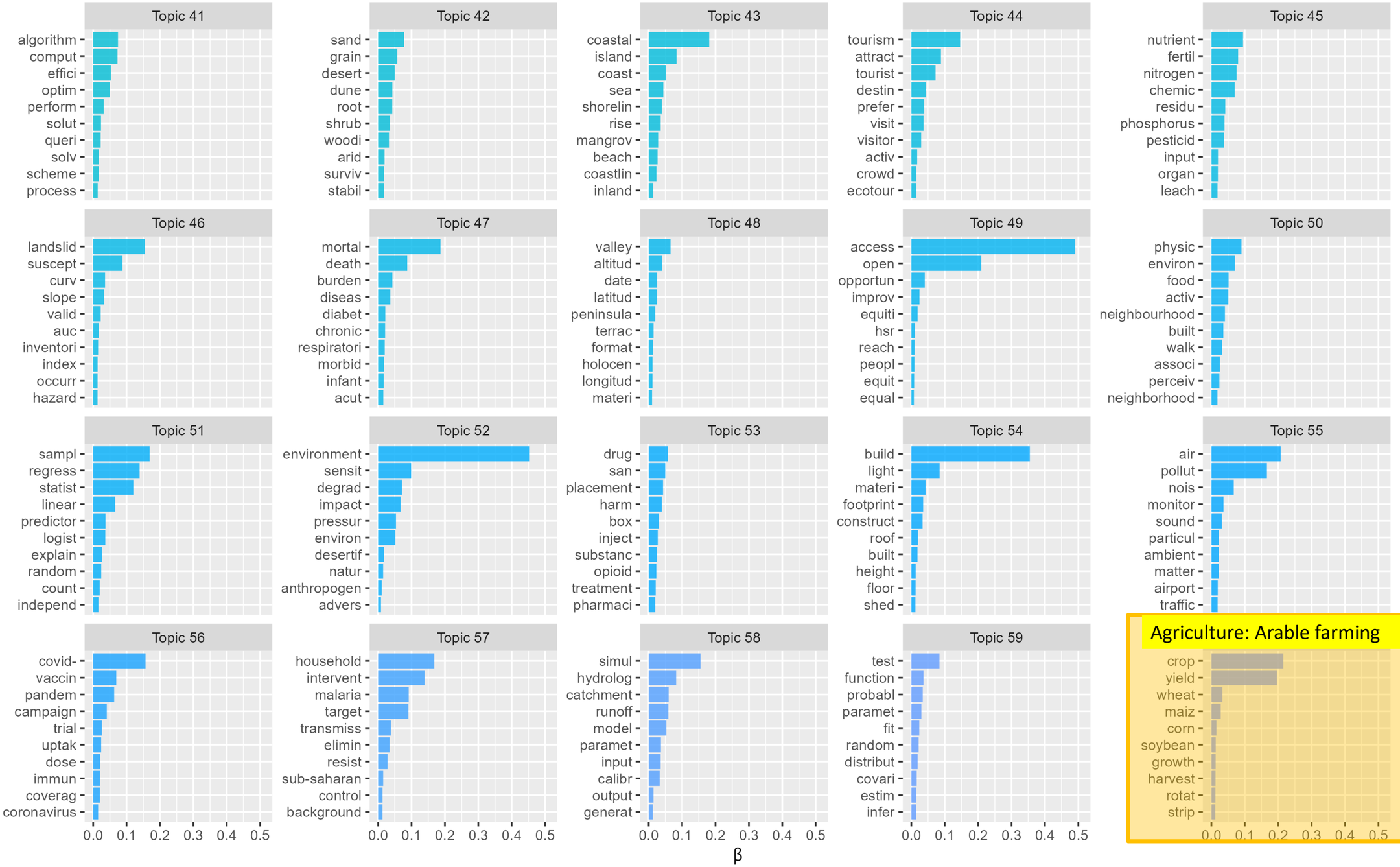


stm::labelTopics

Qualitative topic evaluation

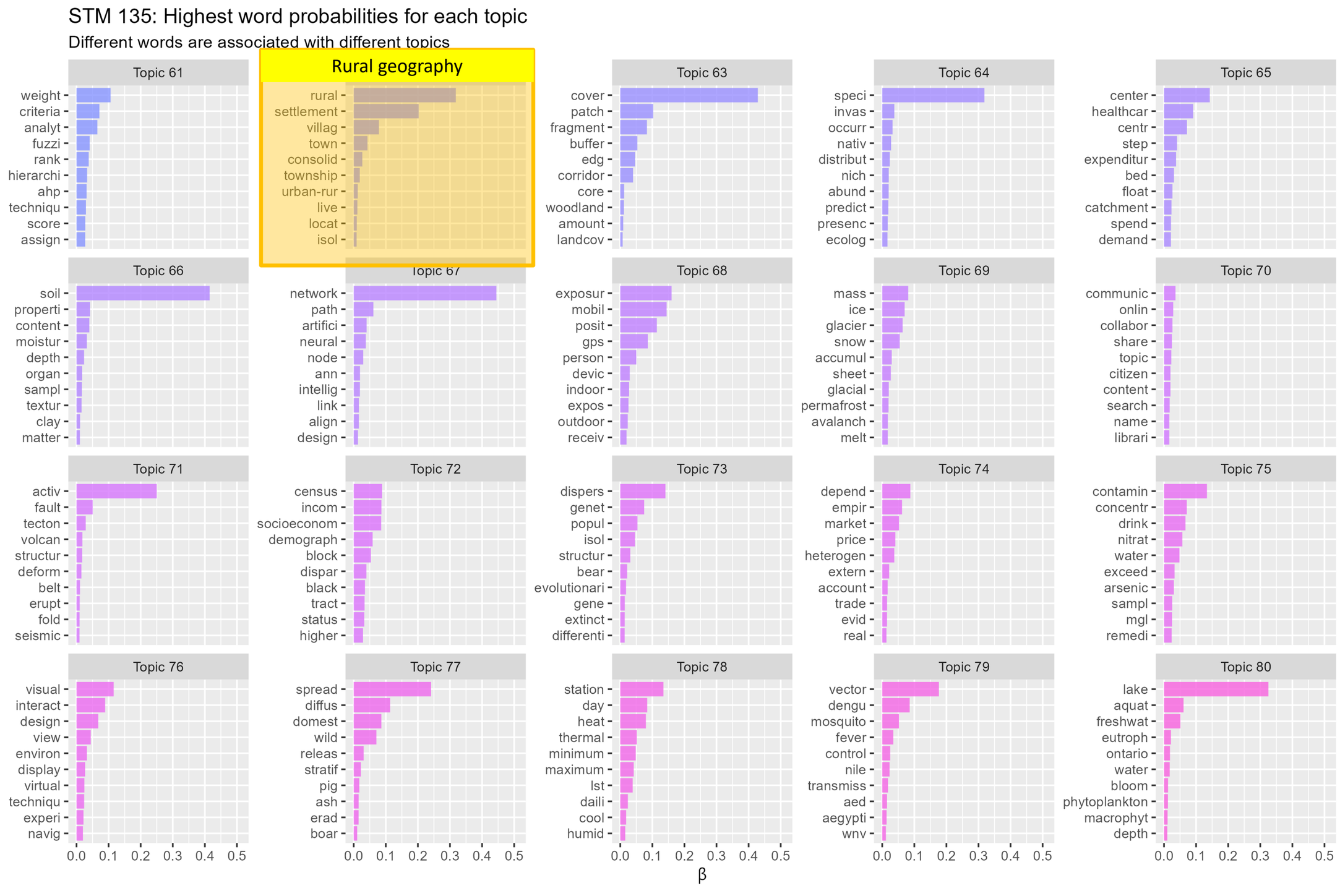
STM 135: Highest word probabilities for each topic

Different words are associated with different topics



stm::labelTopics

Qualitative topic evaluation



stm::labelTopics

Qualitative topic evaluation

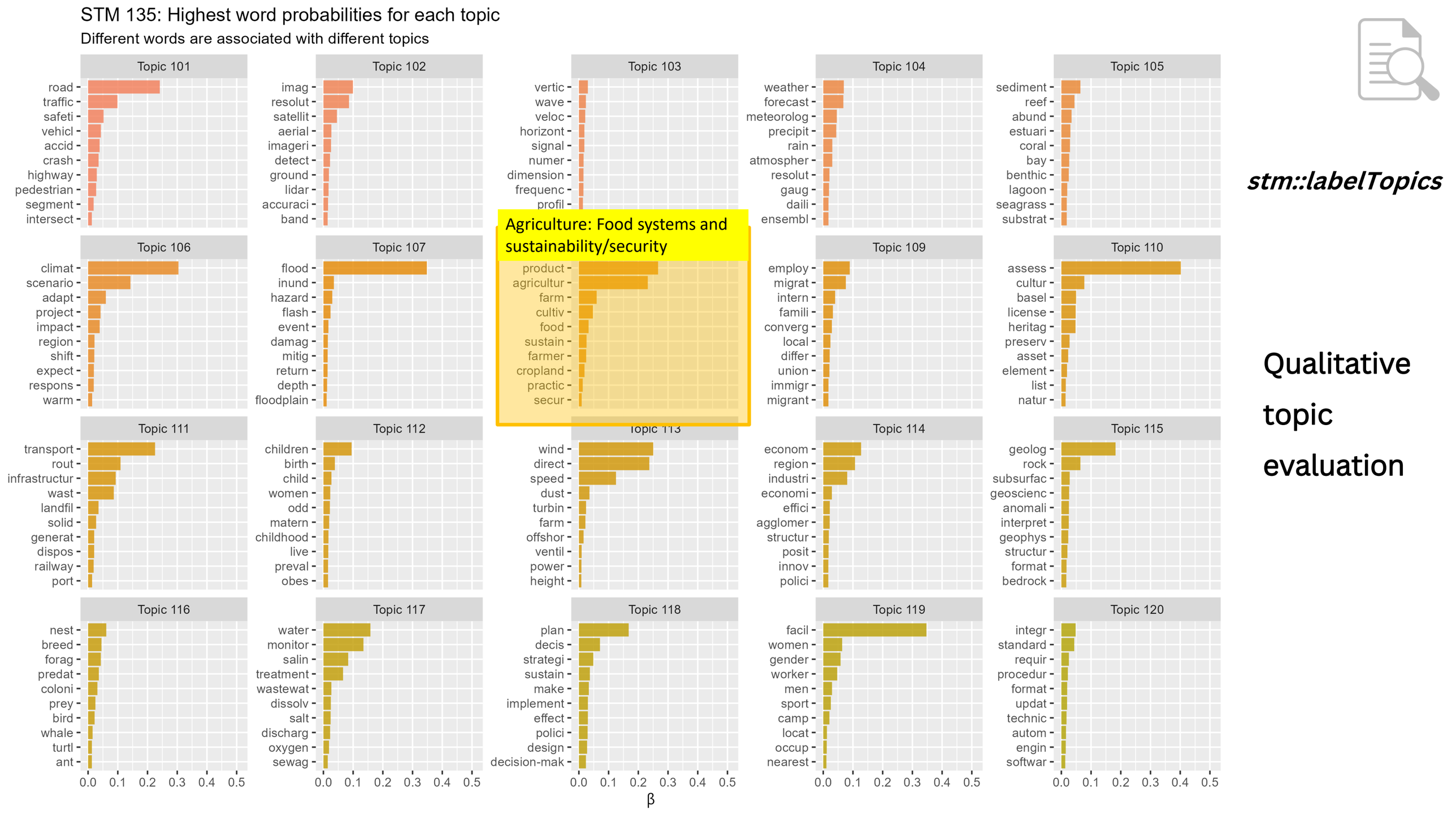
STM 135: Highest word probabilities for each topic

Different words are associated with different topics

Agriculture: Irrigation & water management

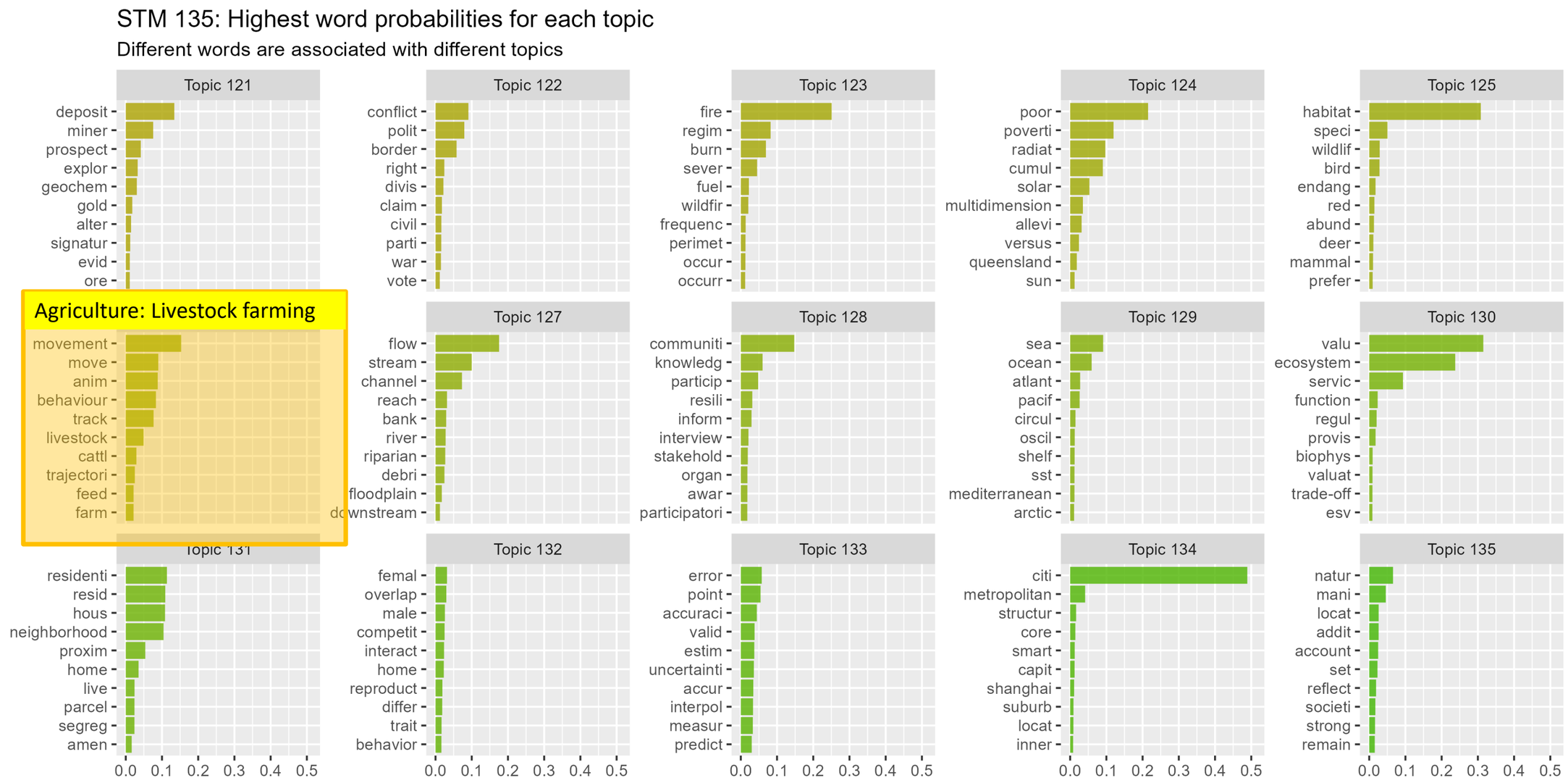
stm::labelTopics

Qualitative topic evaluation



stm::labelTopics

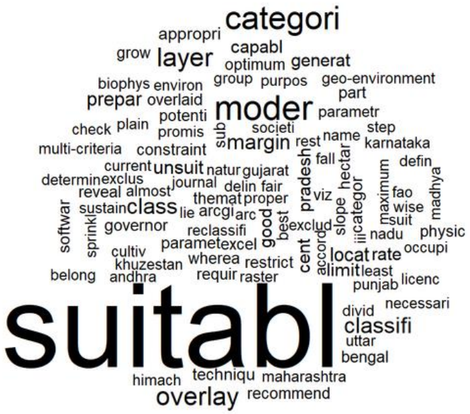
Qualitative topic evaluation

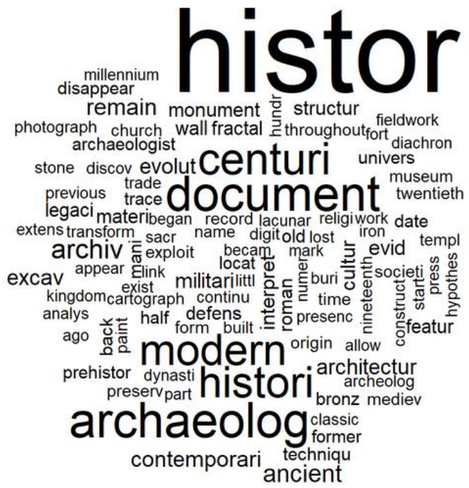


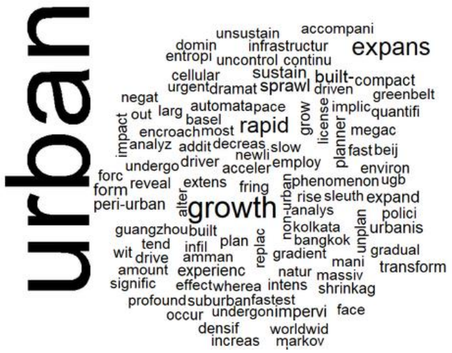
stm::labelTopics

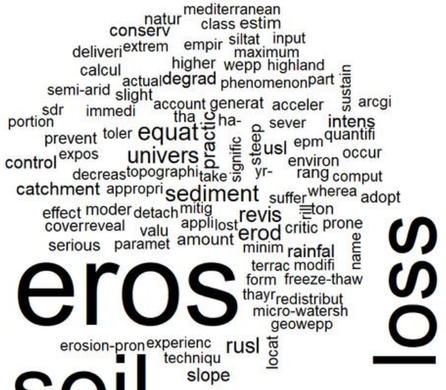
Qualitative topic evaluation

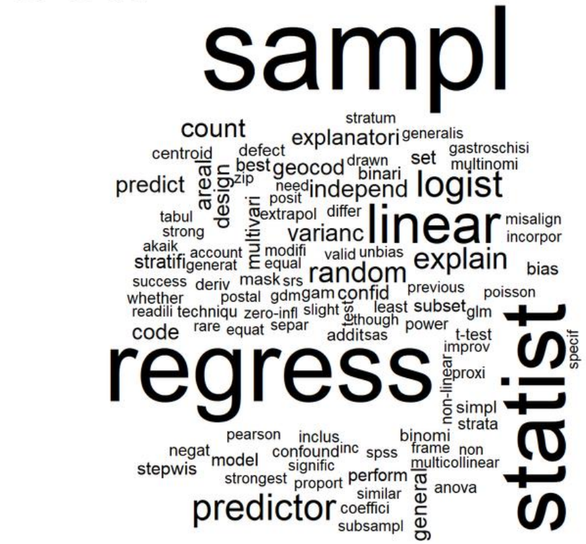
stm::wordCloud

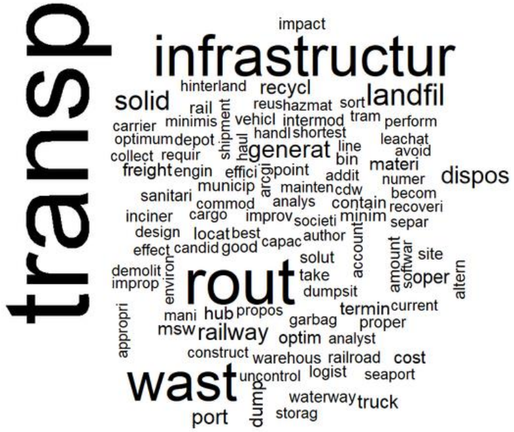


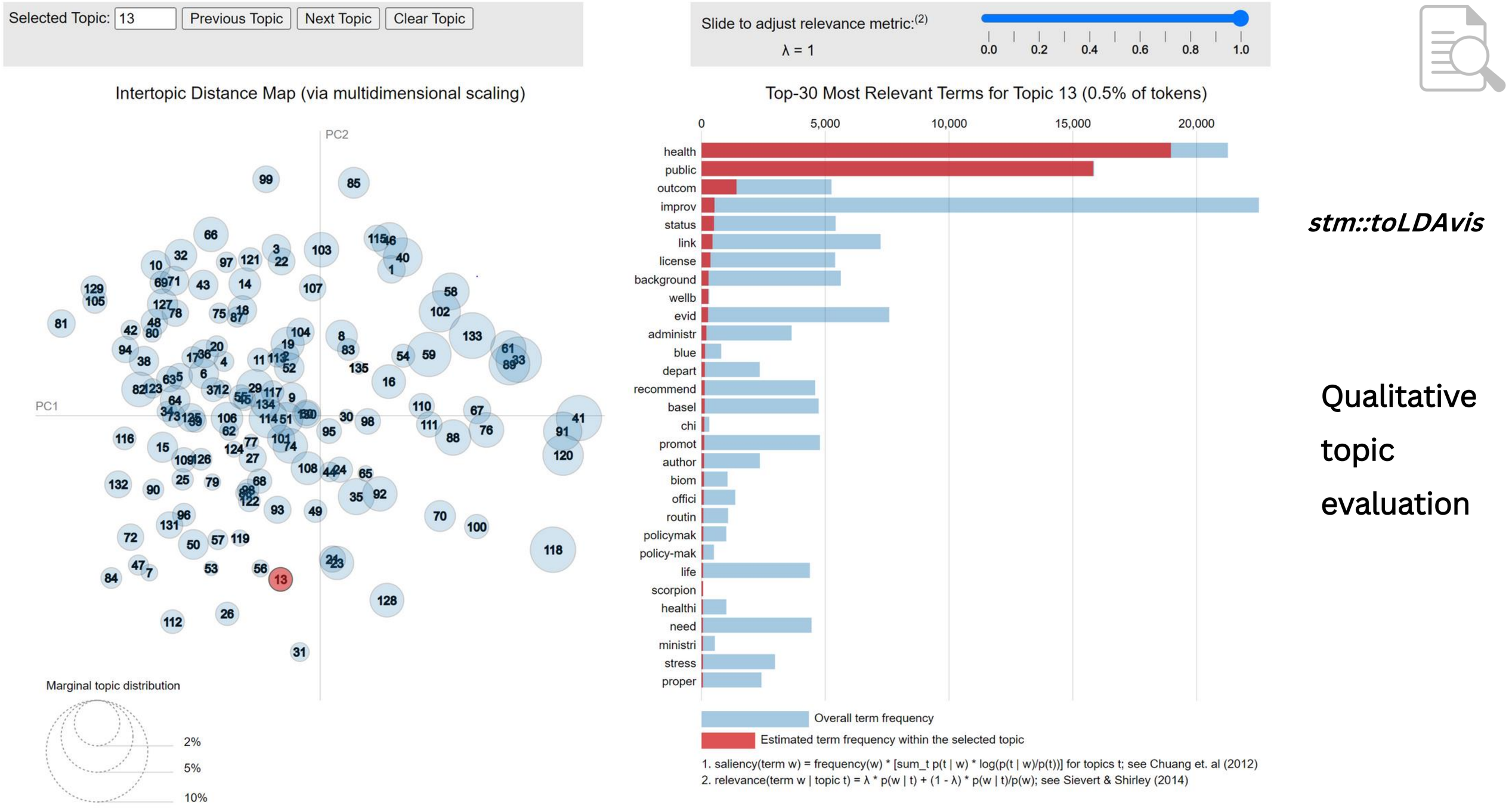












stm::toLDAvis

Qualitative topic evaluation

stm::findThoughts

findThoughts_STM_Final_Topic_60 - Notepad

File Edit Format View Help

Topic 60:
the on in yield gaps (differences between actual yields and the theoretically attainable yields) restricts the of rational strategies to optimize yields and environmental costs. quantifying the yield and the in yield gaps that yields and a narrowing of the yield gap. here, we an analytical to yield to options for closing the yield gap at the . we a yields for 40 and from 87 on-farm experiments in , , from to . the yield was simulated for each - using a hybrid-maize (http://www.hybridmaize.unl.edu/) and weather . we then a and of actual yields to yield gaps at the . the simulated yield at 27 was 15.2 mg ha-1 (8.1–17.6 mg ha-1) in . the on-farm experiments an attainable yield ranging from 8.7 to 16.7 mg ha-1 . during this , the actual maize yield between 4.1 and 11.9 mg ha-1, to the - . farmers' , , 52% of the yield and 77% of the attainable yield. widely amounts of p fertilizer input farmers significantly to regional in yge. soil olsen-p and rainfall were . the that there is to substantially the maize yield in non-optimal p , such as in the . , improvements in regional p strategies, such as at the , to be separately to a for the actual maize yield. © shi, zhang, peng, shi, , liu, , song, cao, cui and cui.
interactions environmental , decisions, and temporal and in corn (zea mays l.) and soybean [glycine max (l.) merr.] yields. the of this are (i) to test whether yield response of corn to n and p and of soybean to p are and temporally , and (ii) to evaluate the profitability of a (vr) n and p fertility strategy over a 5-yr, corn-soybean rotation using this response . a near windom, mn, , was cropped with corn (, , and) and soybean (,). replications of 13 n and p treatments were in a split-plot arrangement of a randomized block design. treatments were at constant in strips the . fertilizer n treatments were 0, 67, 112, 157, and 202 kg ha-1 and p treatments were 0, 56, and 112 kg p2o5 ha-1. the was partitioned into sub-blocks for of yield response. corn and soybean response to these inputs was for each block, each . that of crop response to these inputs is , and that response of corn and soybean to p is temporally in some parts of the , but not . response to n was not temporally . of an ex post profitability that returns over the 5-yr from the vr n and p strategy were $28 ha-1 higher than returns from a uniform strategy. © society of agronomy.
yield gap (yg) is to opportunities for yield increases. is of the most productive soybean in the . in this , soybean is planted after a winter fallow (from on soybean as crop) or after the harvest of a winter crop (from on soybean as crop). options for obtaining higher yields is . the of this are: i) to yg of soybean as or crop, ii) to and environmental with soybean yg , and iii) to the of soybean yg. a farmers' with ~22,500 from to was compiled. water- yield (ywlim) was as the 95th percentile of actual farmers' yield (ya) . yield gap was the between ywlim and ya, expressed as a of ywlim. with yg were evaluated using regression trees. ordinary kriging was to of yg. ywlim were 5095 and 4337 kg ha-1 for and crop, . yg were 28.7 and 33.5% for and crop, . yield gap a of . accounted for 66 and 91% of explained in yg for and crop, . gap closing for crop was with planting and maize as crop. gap closing for crop was with foliar fungicide , p fertilization, and planting. crop yg was auto-, whereas no auto- was for crop. the structure of crop was by an exponential , with 81% of explained by the structure and a maximum of auto- of 120 km. this is with the auto- of explaining yg in crop. our approximation the of the magnitude, explaining , and dependence of soybean yg in of the most productive in the . gaps are to those in , there are opportunities for yield improvements. © .
yield (yp) and yield gap (yg) in intensive potato (solanum tuberosum l.) production is to meet food demand with the . evaluating yield gap is a strong to maximum production when are in the best . a estimation of yield gap and yield potato in is lacking. the yield gap atlas (gyga) protocol was to yield of potato in . this protocol is on the climatic (czs) and the reference weather stations (rws) buffer , soil in each buffer . thirty- rws buffer in potato were , and potato in the rws buffer covered 83% of the potato harvest . to the , the yp was 67.3 t ha-1 and actual yield (ya) was 30 t ha-1. , the tuber yield gap was 37.3 t ha-1. these that the potato producers 45% of the yield in . farmers 5 tons of potato from about 164,000 ha. if they only 80% of yp (53.8 t ha-1), amount of potato production be 8.8 tones. as , they 5.2 tons tuber yield of potato in 97,000 ha cultivation . , with closing yield gap and potato production, it is to decrease potato lands
of of gaps between yield and producer yields has been restricted to . in the , we a novel for of yield gaps over agricultural with diversity in climate and soils. this was to quantify and explain yield gaps in rainfed and irrigated soybean in the - (nc) , which accounts for about third of soybean production. on yield and were from 3568 producer over crop seasons and grouped into 10 extrapolation domains (teds) to their soil, climate, and water regime. yield was using a of crop modeling and functions for water productivity and against producer yields derived from the yield in each ted-. yield gaps were as the between yield and producer yield. explanatory for yield gaps were by practices that were concordantly with - and -yield . × ted interactions were then evaluated to elucidate the underlying of yield gaps. the ted accounted for about half of the regional in producer yield the nc . the 10 teds, soybean yield ranged from 3.3 to 5.3 mg ha-1 for rainfed and from 5.3 to 5.6 mg ha-1 for irrigated . producer yields in each ted were similar (±12%) to the yield . yield gap, as of yield , was in rainfed (: 15–28%) than in irrigated (: 11–16%) soybean. upscaled to the nc , yield was 4.8 mg ha-1 (rainfed) and 5.7 mg ha-1 (irrigated), with a respective yield gap of 22 and 13% of yield . sowing date, tillage, and in-season foliar fungicide and/or insecticide were as explanatory for yield in half or of the 10 teds. , the degree to which these producer yield teds. of in-season weather interpret × ted interactions. for , yield to in sowing date was in teds with water during the pod-setting . the the strength of producer with a to yield gaps, explaining these gaps, and the biophysical drivers yield responses to crop
of yield trials adjustment of cultivar for , improving the statistical precision of yield estimation. while the efficiency of has been frequently in yield trials, its to - lolium spp. forage yield trials has not been . the of this was to evaluate the , nearest-neighbor (nna), and error (ce) for their ability to account for in 138 lolium spp. forage yield trials. this was performed on from and 11 yr (-) using randomized block design (rcbd) trials by the department of agriculture, food and marine (dafm) in . the efficiencies of , nna, and ce with rcbd were 129, 143, and 193% for by trial ×

Ln 1, Col 1 100% Windows (CRLF) UTF-8

The topic of this conversation is related to forest management and conservation. The first text describes a study on reforestation and deforestation trends in a rural mountainous area in Lika, Croatia, and discusses the potential benefits of reforestation efforts for socio-economic planning. The second text focuses on the relationship between forest loss and fire-induced damage in a particular region, providing insights on the types of forests most affected by these phenomena. The third text examines the effects of deforestation and logging on forest fragmentation and edge creation in the Amazon rainforest, highlighting the ecological repercussions of these activities.

The broad topic of this text is multi-criteria decision-making (MCDM) techniques, which are used to evaluate and rank feasible alternatives based on multiple conflicting and noncommensurate criteria. The text describes various MCDM techniques such as the technique for order of preference by similarity to ideal solution (TOPSIS), Borda's rule, fuzzy analytic hierarchy process (FAHP), and linguistic quantifier-guided weighted averaging (OWA) procedure. These techniques are applied in various decision-making scenarios, including wind farm implantation, hospital planning, solar photovoltaic farm deployment, land suitability evaluation, and solar thermoelectric power plant site selection.

The overall broad topic of these texts is economics, with specific topics including mergers and competition, service provision in government, remote sensing and marketing, gambling industries, government dissemination policies, and fiscal decentralization.

The overall broad topic appears to be healthcare and medical research, with a focus on various aspects such as acute care in patients' homes, transportation of trauma patients, delay times in surgical care, and adherence to surgical infection prevention guidelines.

The overall broad topic of these texts is topological relations and their applications in various fields such as object-oriented structure, generalization, query, reasoning, and shape simplification. The texts discuss the definitions and properties of topological relations between objects, and how they can be represented using matrices, sets, and formal languages. They also explore the challenges and limitations of existing approaches and propose new methods to address them, such as hierarchical invariants, multi-modeling, and shape decomposition. Additionally, the texts highlight the importance of constraints and operations in guiding and governing generalization, and the role of vagueness in modeling complex phenomena.

stm::findThoughts **+ ChatGPT-4**

The broad topic of these passages seems to be sociospatial theory, which is concerned with the relationship between society and space and how they shape each other. Specifically, the passages explore various aspects and perspectives of sociospatial theory, including its conceptualization, its relationship to politics and power, and its applications in different contexts such as globalization, post-representational cartography, and postsecular theory.

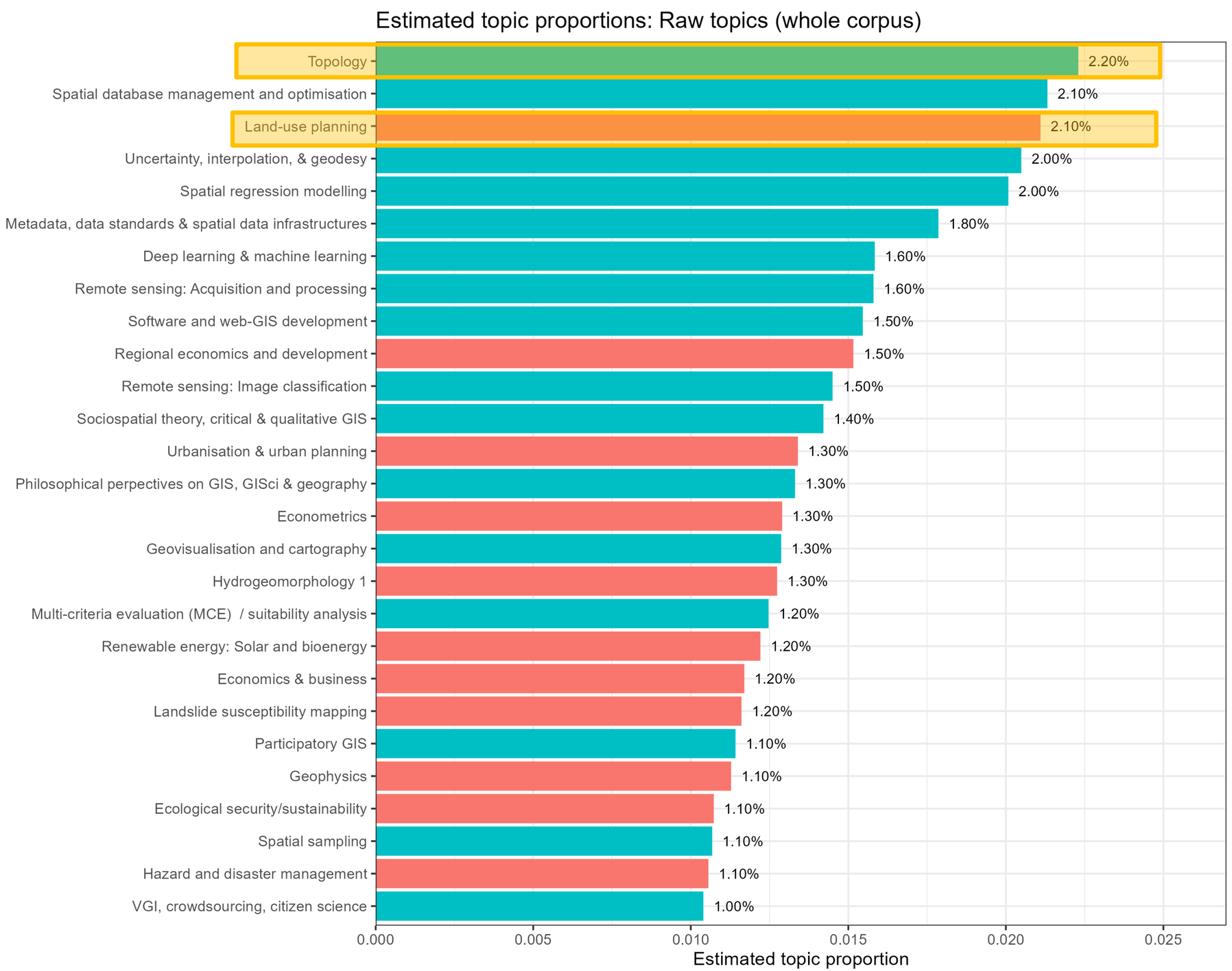
STM 135

GIS&T Body of Knowledge

http://gistbok.ucgis.org/

Maintains a list of topics in GIS

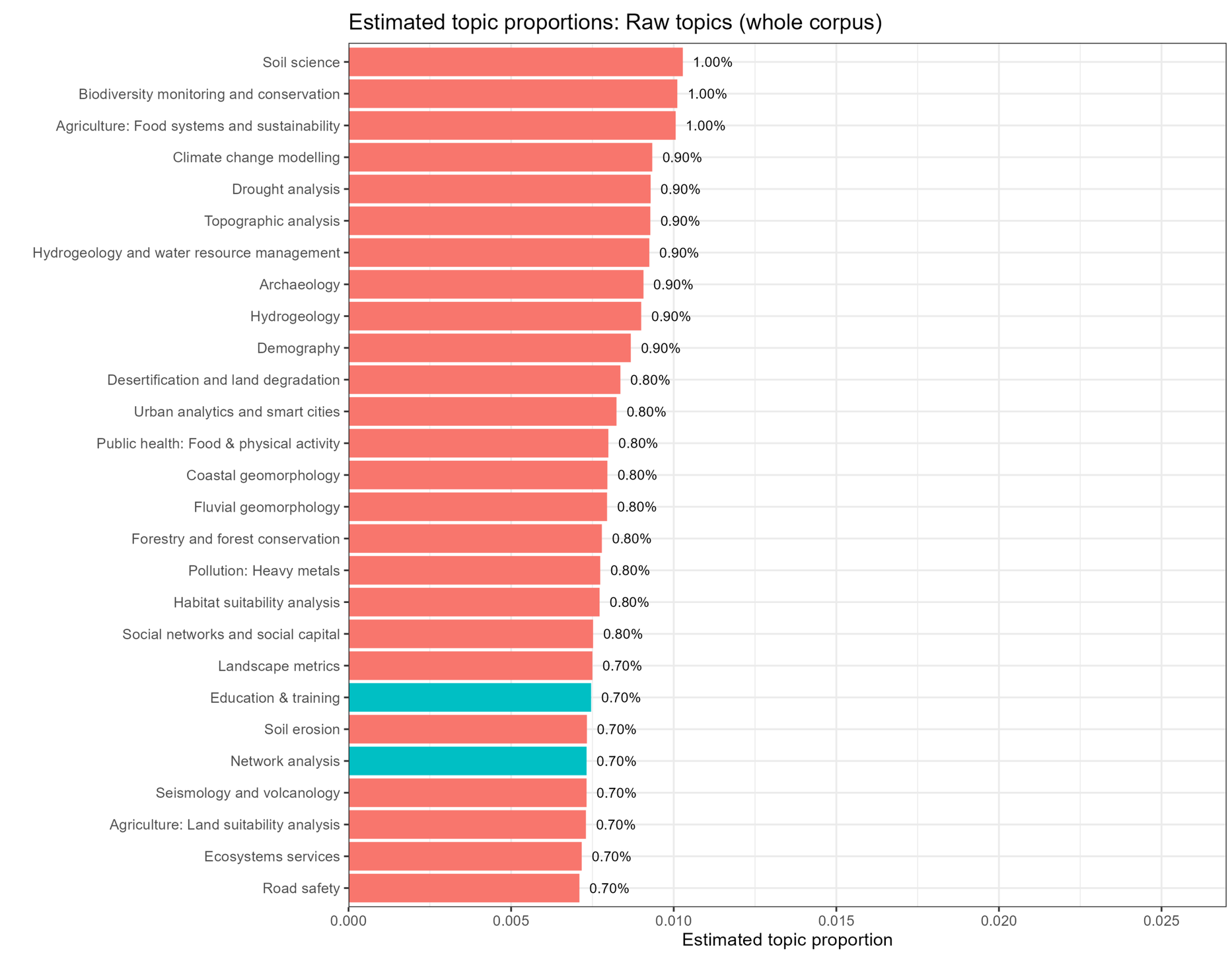
Results



Whole corpus
(all years)

Domain application
GISci

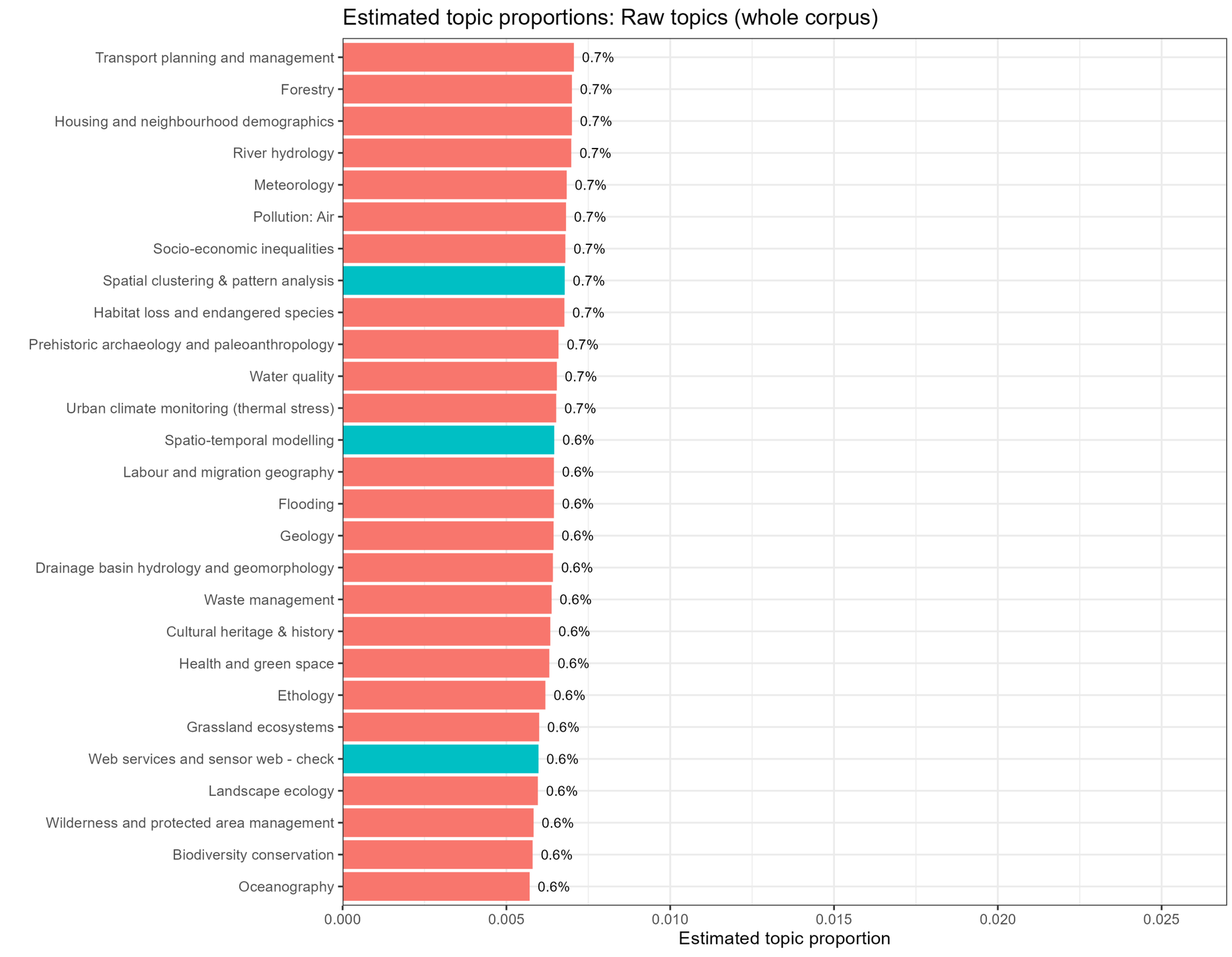
Raw topics
Level 1
(135 topics)



Whole corpus
(all years)

Domain application
GISci

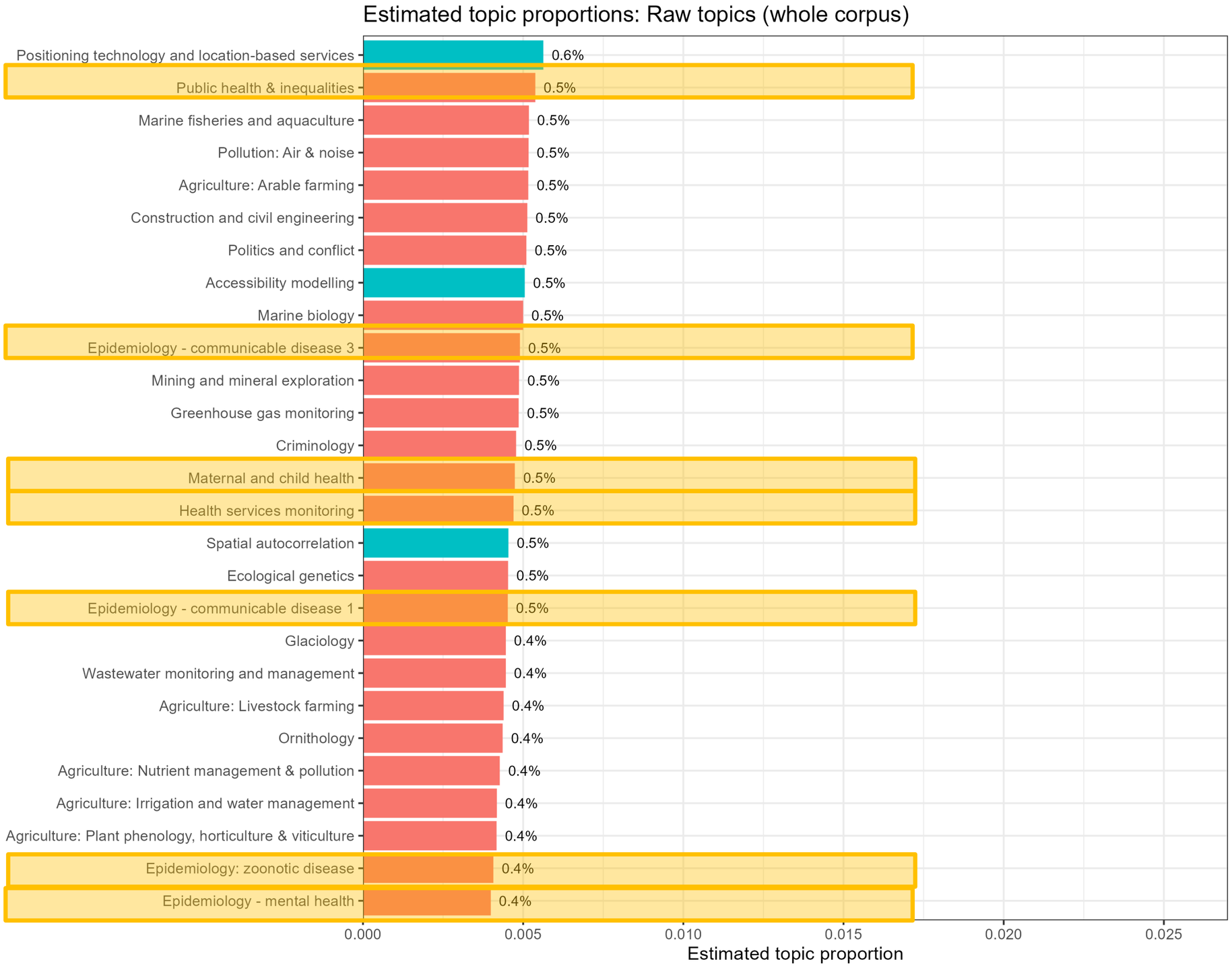
Raw topics
Level 1



Whole corpus
(all years)

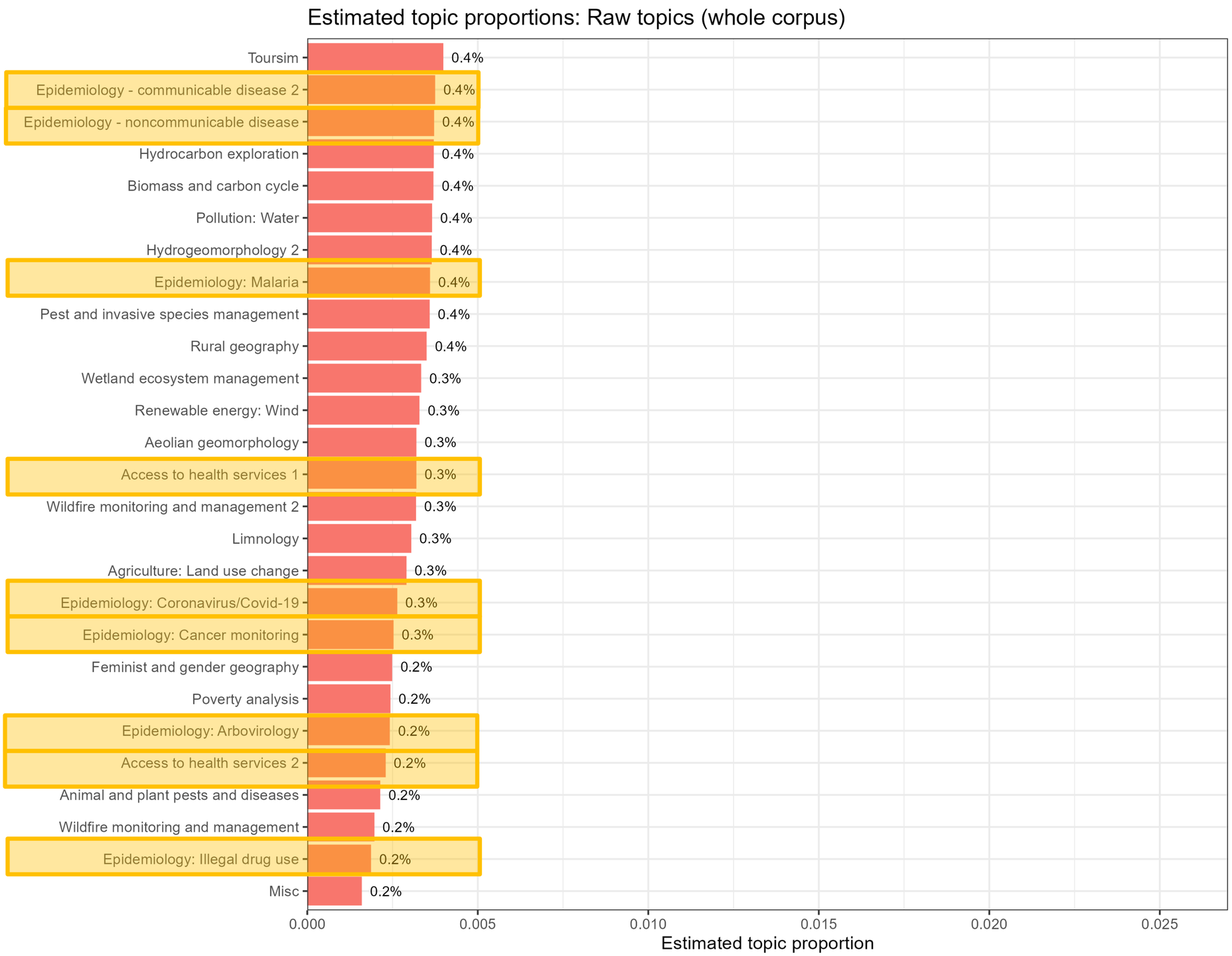
Domain application
GISci

Raw topics
Level 1



Whole corpus
(all years)

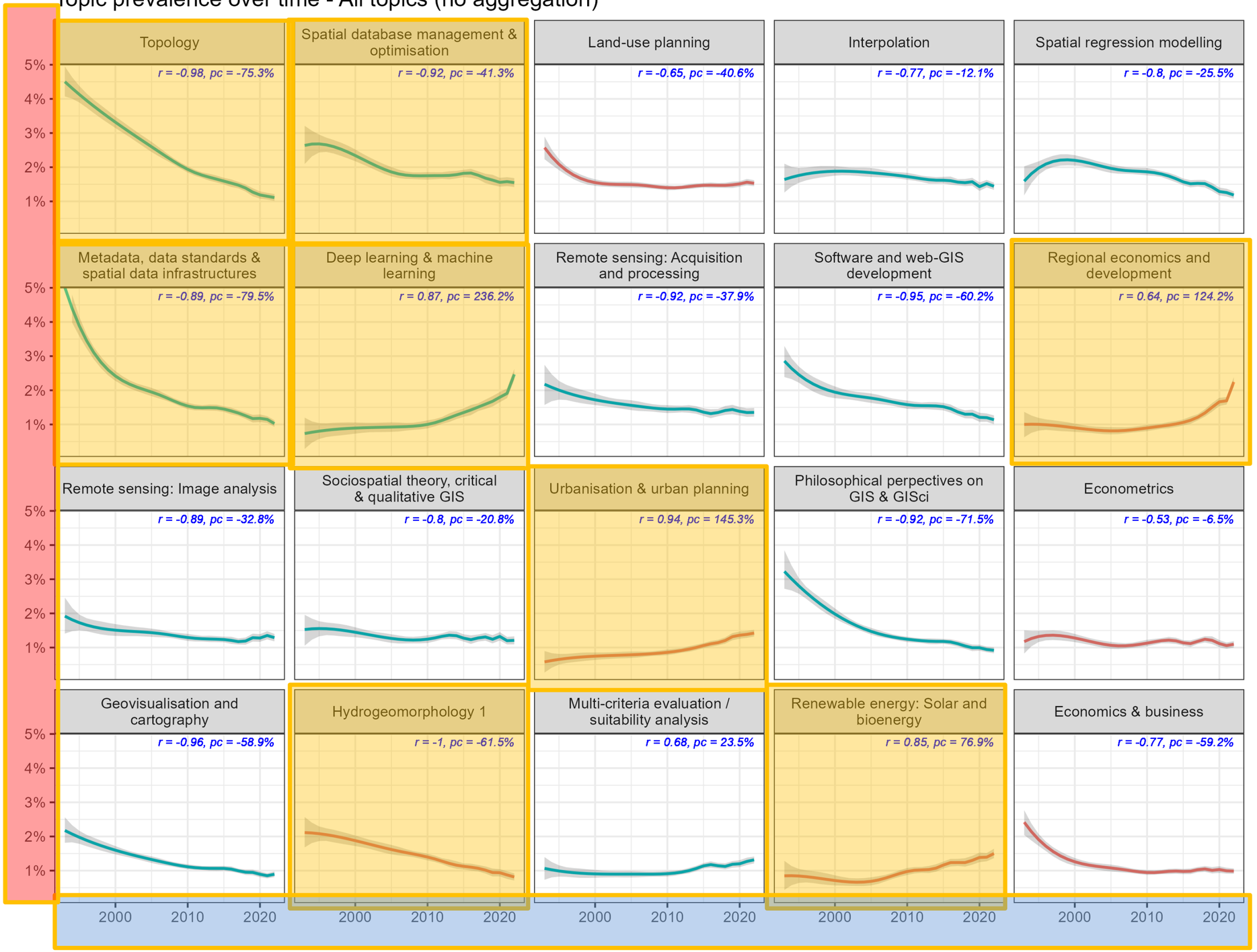
Raw topics
Level 1



Whole corpus
(all years)

Domain application

Raw topics
Level 1



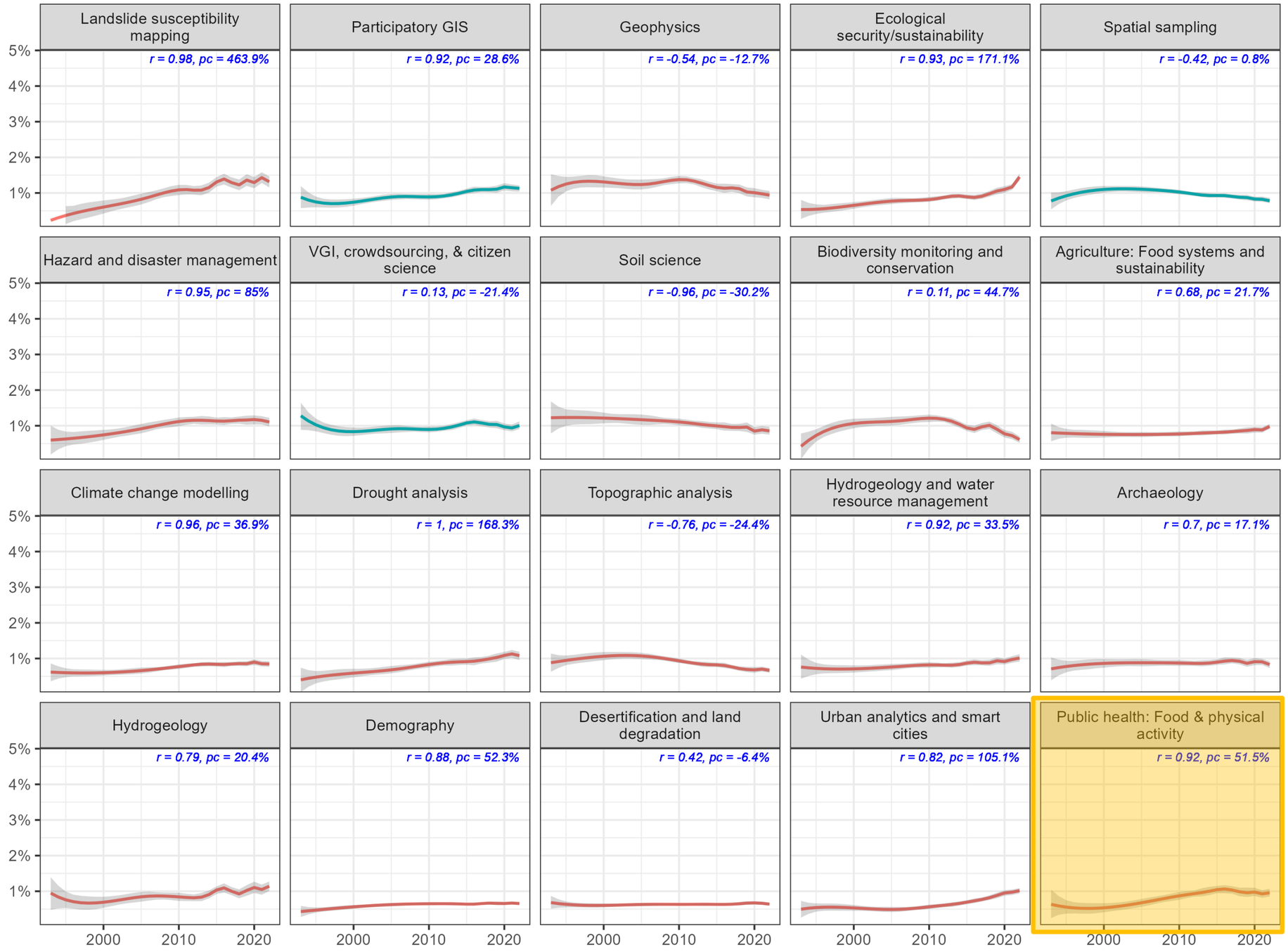
Topics over time

Domain application

GISci

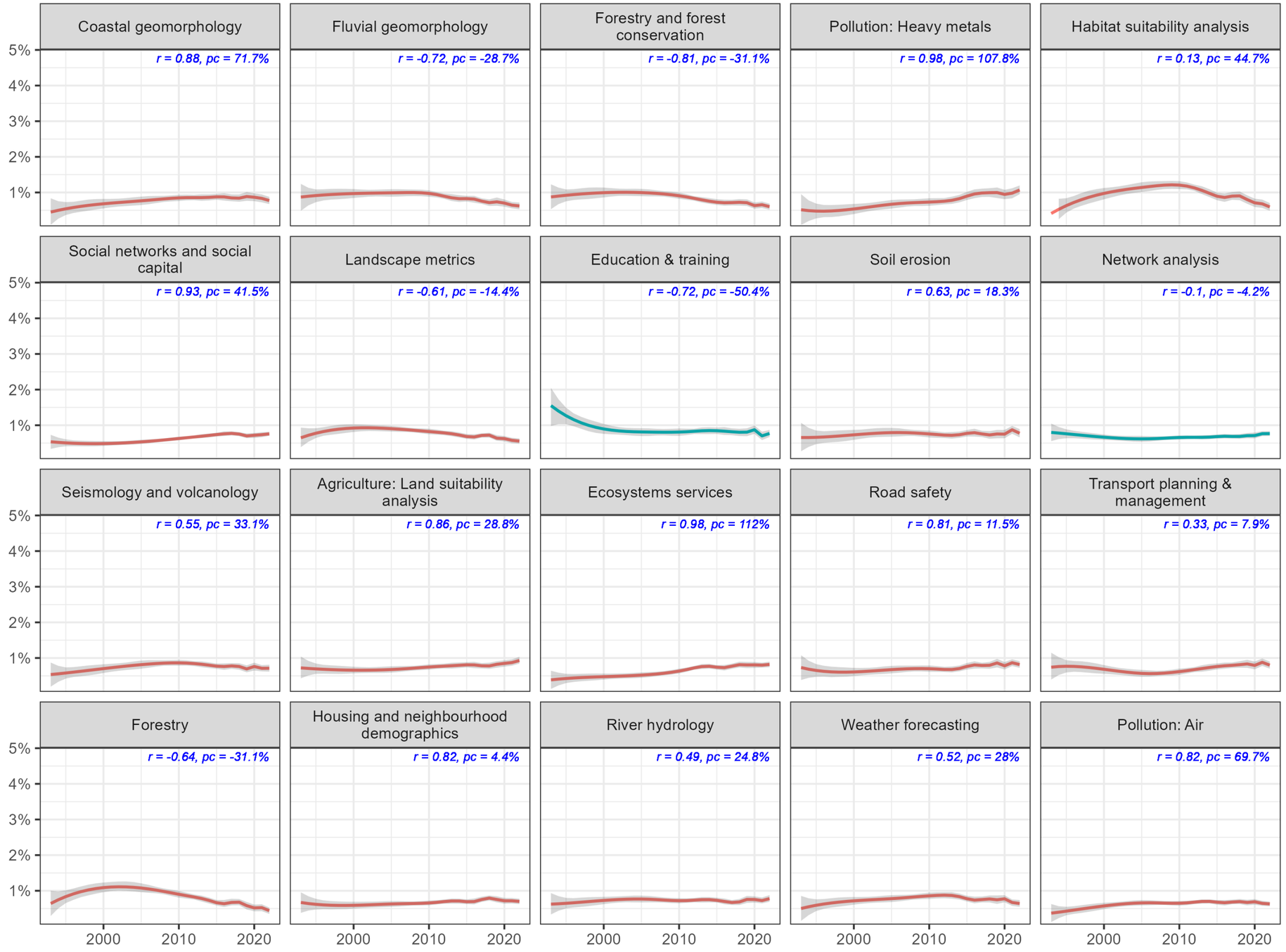
Raw topics

Level 1



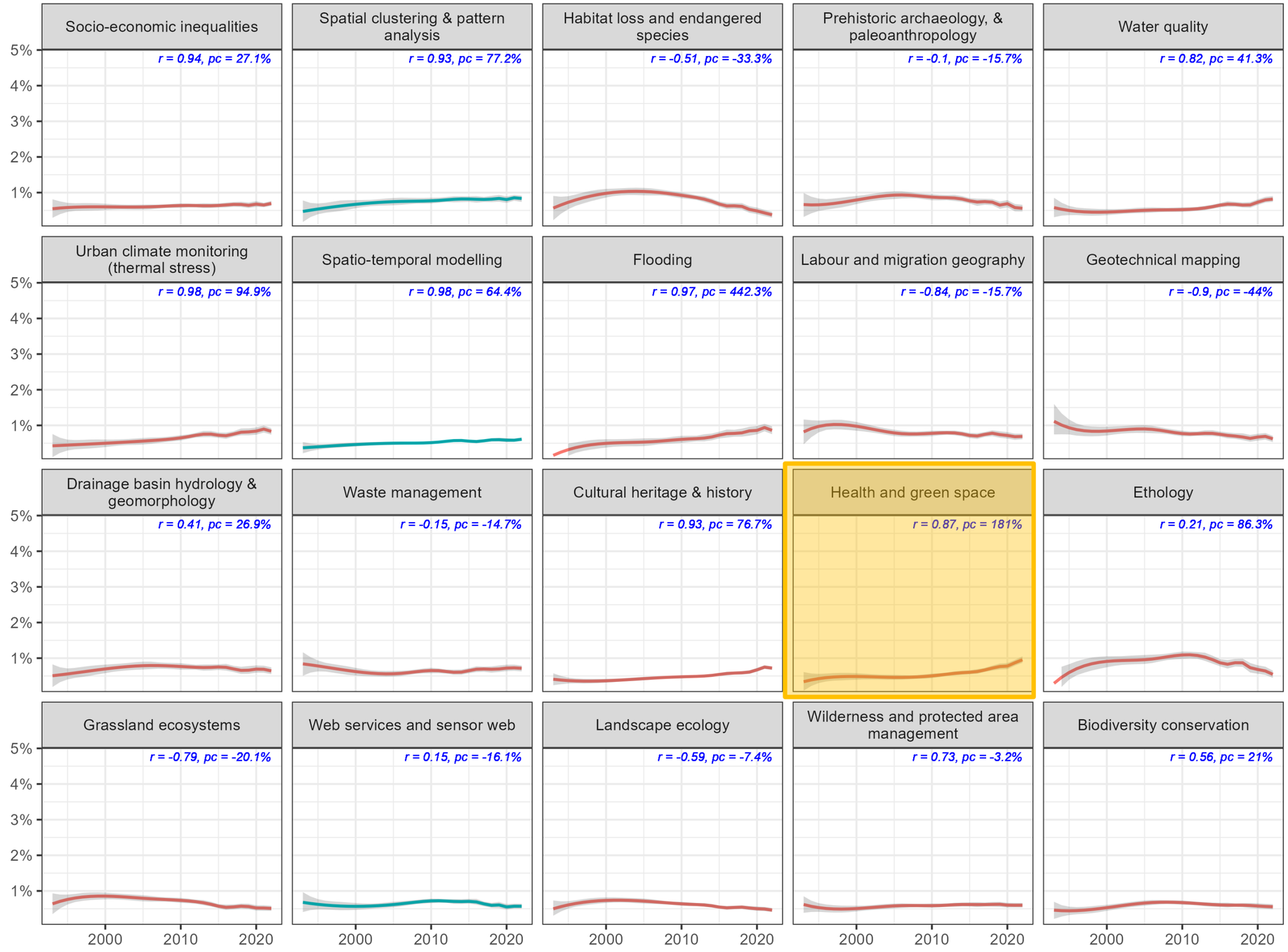
Raw topics

Level 1



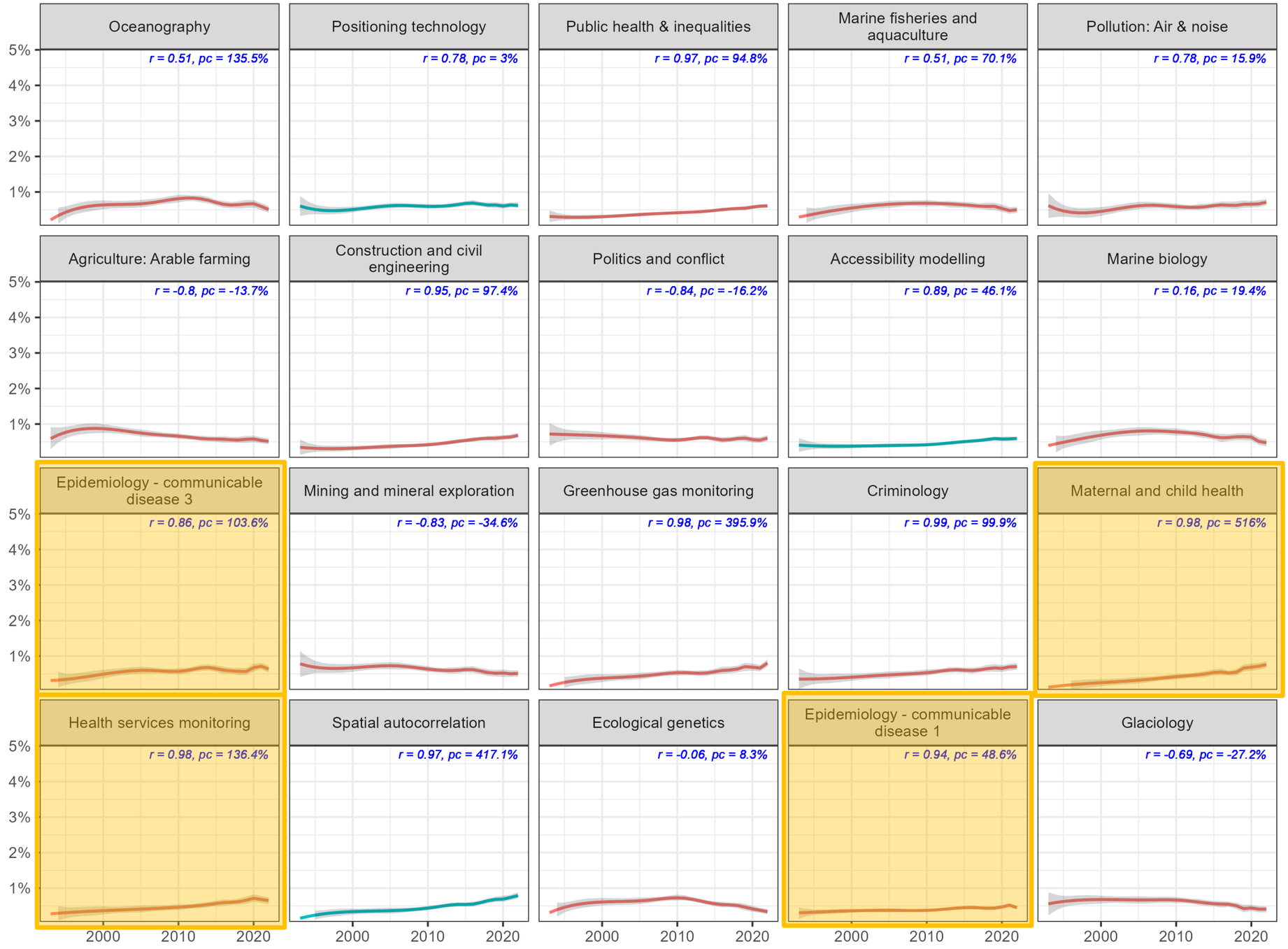
Topics over time

Raw topics
Level 1



Topics over time

Raw topics
Level 1



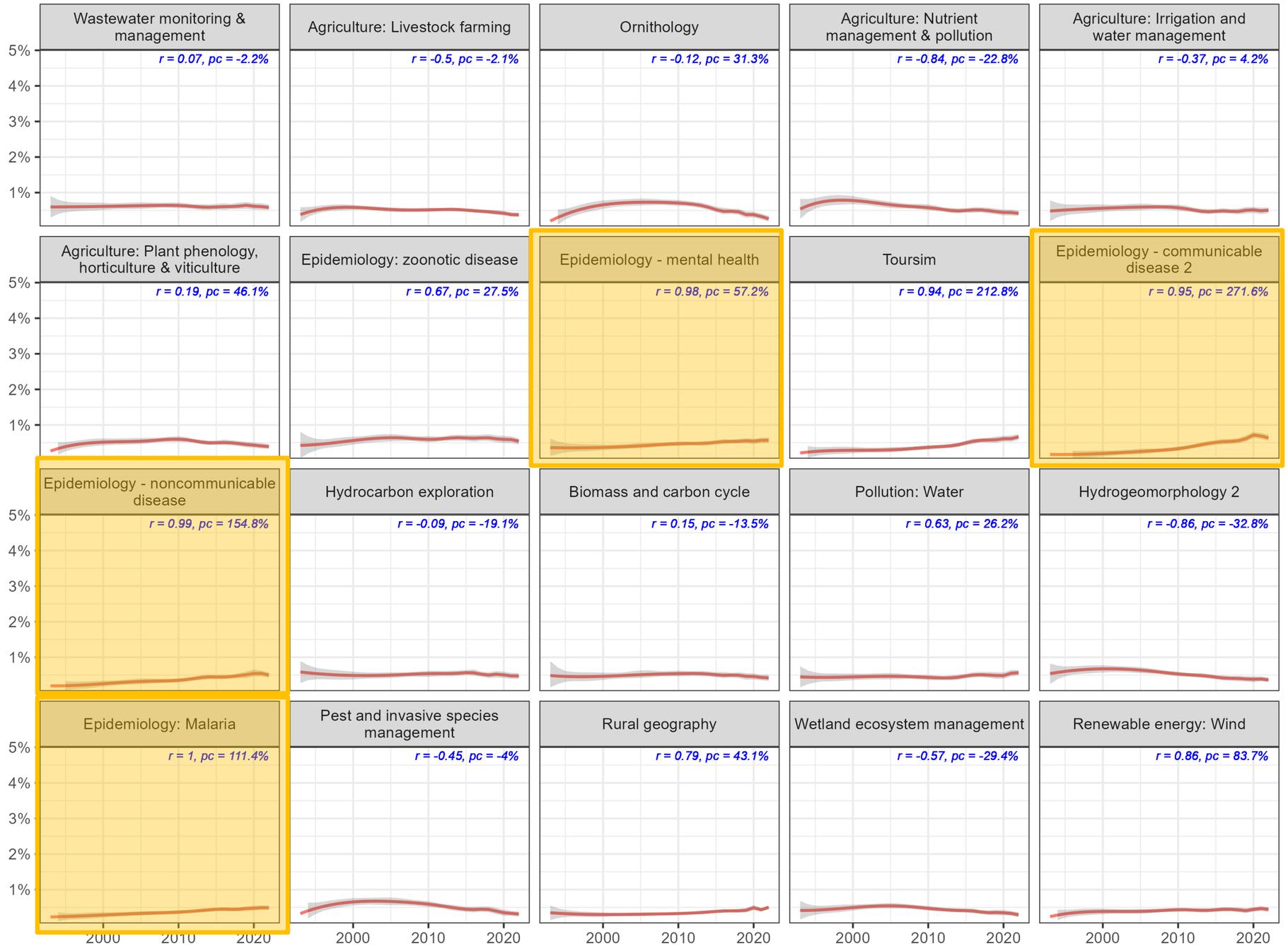
Topics over time

Domain application

GISci

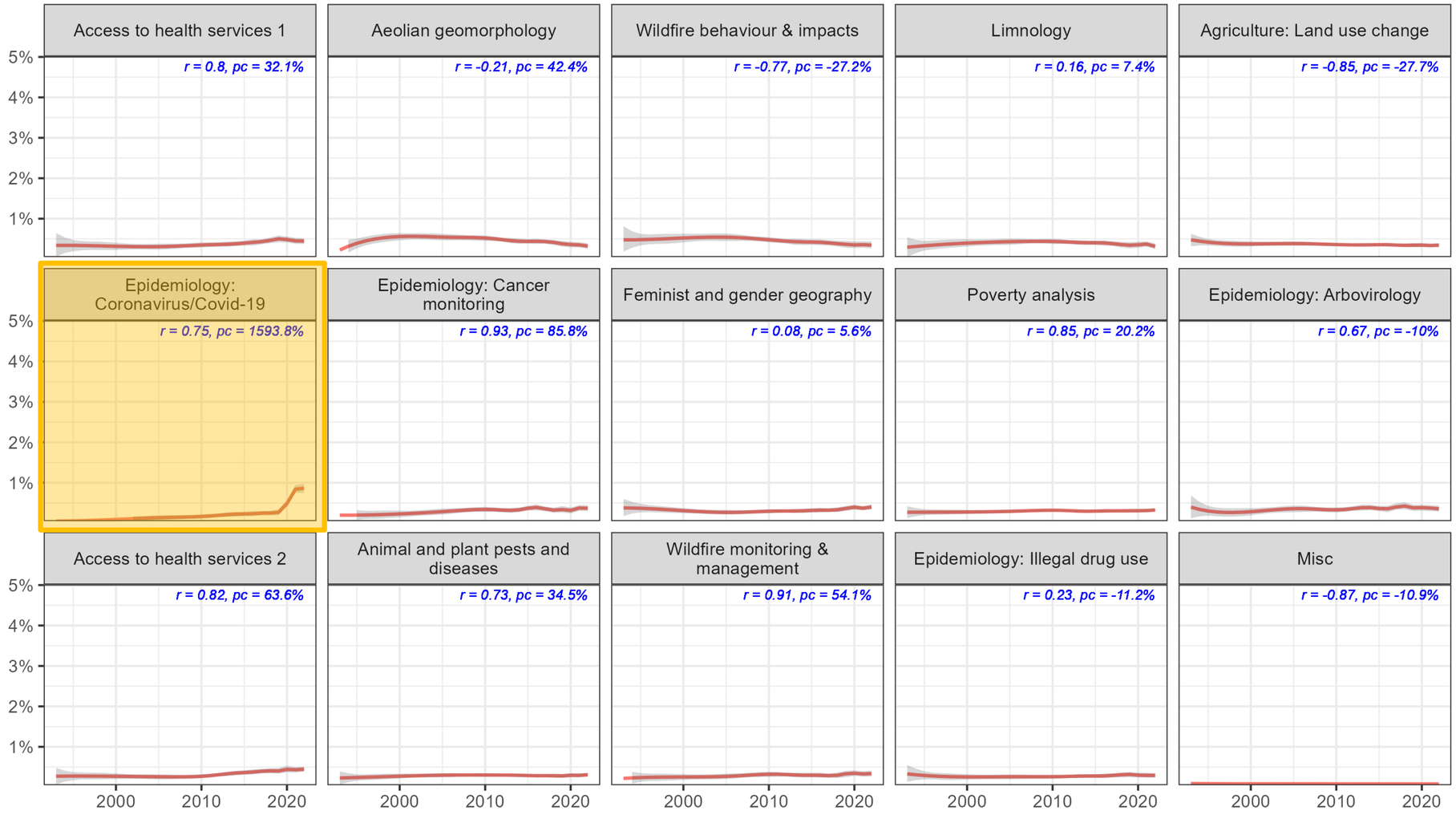
Raw topics

Level 1



Topics over time

Raw topics
Level 1



Topics over time

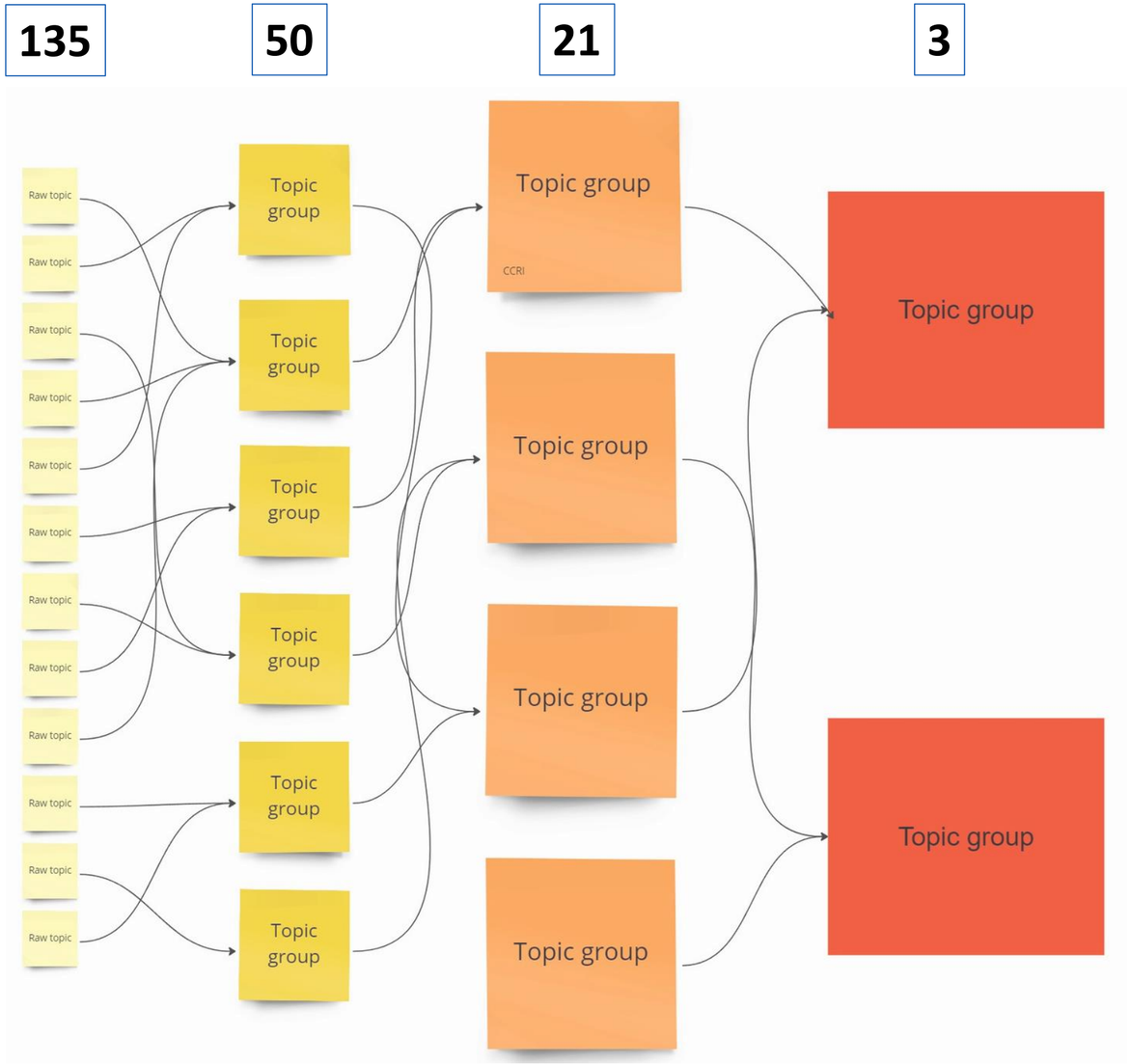
Domain application

GISci

Raw topics

Level 1

STM 135 - Aggregating topics



4 levels:

- Raw topics (**Level 1**) to highly aggregated topic groups (**Level 4**)

Using:

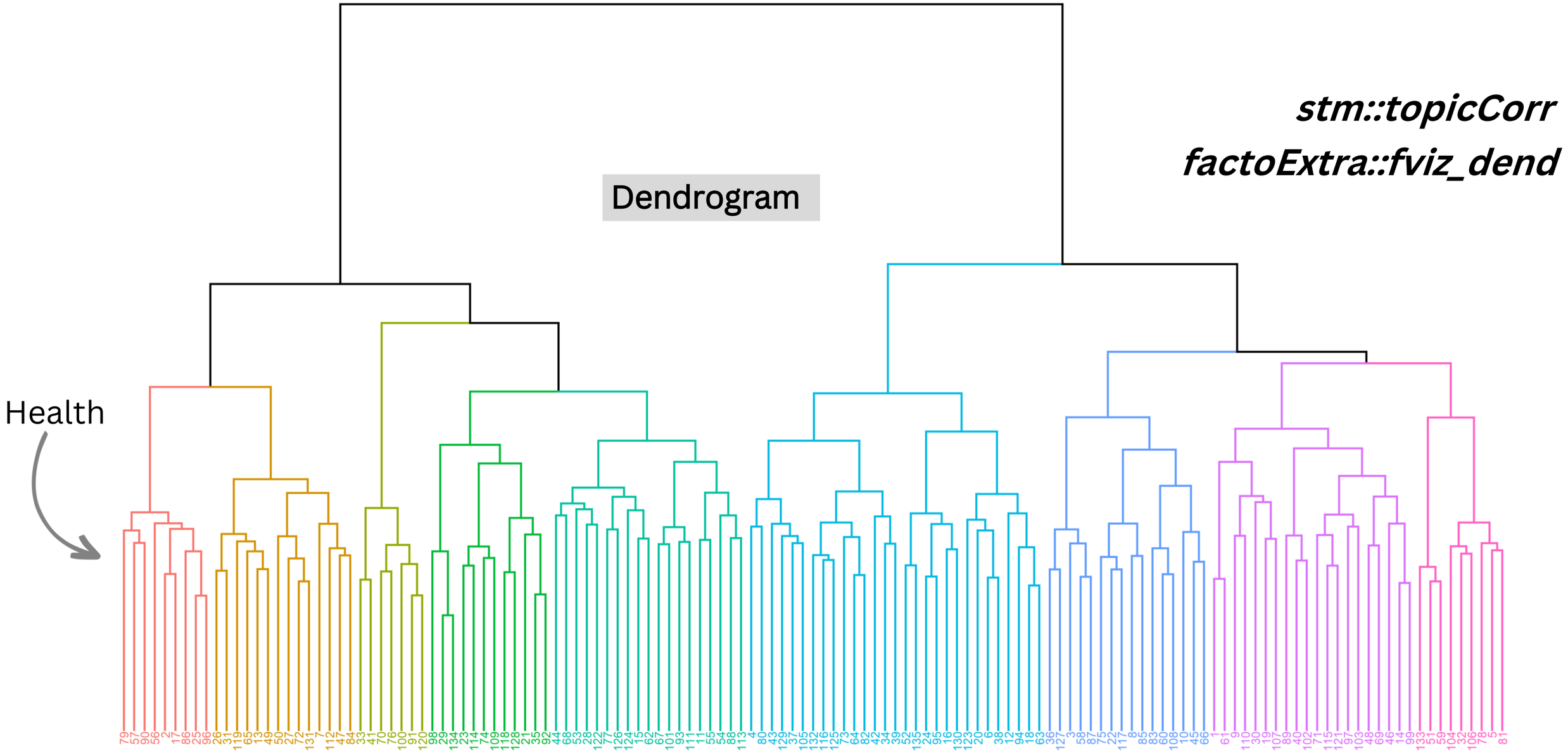
- Domain knowledge
- Statistics (correlations)
- Literature

Aggregating topics: topic correlations

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20	Topic 21	Topic 22	Topic
Topic 1	1	-0.04372	0.02149	0	-0.06612	0.01719	-0.0205	0.06309	-0.04363	-0.01353	-0.02111	0.01896	-0.04969	0.05224	-0.01584	0.04402	-0.02831	-0.01358	-0.01484	0	-0.03806	0.0566	-0.03
Topic 2	-0.04372	1	-0.04033	-0.02821	0.10908	-0.03504	0.08514	-0.04832	-0.02803	0	-0.02331	-0.02256	0.05873	-0.05672	0.06435	-0.0204	0.23075	-0.02551	-0.02324	-0.02835	-0.01444	0	-0.02
Topic 3	0.02149	-0.04033	1	0.01012	-0.02784	0.01846	-0.0155	0.04509	-0.02173	-0.01479	-0.01507	0.01892	-0.05175	0.09382	-0.04739	0	-0.0289	0	0	0	-0.02887	-0.01814	-0.03
Topic 4	0	-0.02821	0.01012	1	-0.01659	0.03261	-0.01427	0.03812	0	0	0	0.02989	-0.03603	0.01476	0	0.0491	-0.02598	0.07077	-0.02306	0.02752	-0.02562	0	-0.
Topic 5	-0.06612	0.10908	-0.02784	-0.01659	1	-0.03475	0	0	0	-0.02287	0.01652	0	-0.03664	-0.04256	0.02493	0.01132	0.07845	0	0	-0.01243	-0.05015	-0.01833	-0.0
Topic 6	0.01719	-0.03504	0.01846	0.03261	-0.03475	1	-0.02288	-0.03474	0	-0.03274	0	0.05623	-0.05915	0.04591	0.01128	0.01657	-0.02008	0.12345	-0.0336	0.162	-0.03797	-0.03375	-0.01
Topic 7	-0.0205	0.08514	-0.0155	-0.01427	0	-0.02288	1	-0.0224	-0.01733	0	0	-0.01805	0.06876	-0.03026	0.02445	-0.01901	0.07195	-0.02402	-0.01592	-0.0134	0	0	-0.01
Topic 8	0.06309	-0.04832	0.04509	0.03812	0	-0.03474	-0.0224	1	-0.03002	0.03764	-0.01279	0.01047	-0.06248	0.02504	-0.05956	0.05008	-0.04531	-0.02691	-0.0143	-0.01575	-0.04765	0.13539	-0.03
Topic 9	-0.04363	-0.02803	-0.02173	0	0	0	-0.01733	-0.03002	1	-0.02919	-0.02993	0.01495	-0.03642	0	0	-0.02052	-0.03674	0.07217	0	-0.02592	0	-0.03426	
Topic 10	-0.01353	0	-0.01479	0	-0.02287	-0.03274	0	0.03764	-0.02919	1	0.03887	0	0	-0.02924	-0.02544	0.02085	0	-0.031	-0.02694	0	-0.02591	0.0907	-0.04
Topic 11	-0.02111	-0.02331	-0.01507	0	0.01652	0	0	-0.01279	-0.02993	0.03887	1	0	-0.02343	-0.03752	-0.01948	0	0	-0.02921	-0.02911	0.06049	-0.02156	0	
Topic 12	0.01896	-0.02256	0.01892	0.02989	0	0.05623	-0.01805	0.01047	0.01495	0	0	1	-0.03707	0	0	0.08361	-0.01692	0.08006	-0.031	0.02484	-0.02003	0	0.02
Topic 13	-0.04969	0.05873	-0.05175	-0.03603	-0.03664	-0.05915	0.06876	-0.06248	-0.03642	0	-0.02343	-0.03707	1	-0.1037	0.10989	-0.02325	0.0888	-0.05918	-0.02678	-0.04556	0.05906	0	0.09
Topic 14	0.05224	-0.05672	0.09382	0.01476	-0.04256	0.04591	-0.03026	0.02504	0	-0.02924	-0.03752	0	-0.1037	1	-0.05737	0	-0.05324	0.02259	0.01696	0	-0.05609	-0.03607	-0.09
Topic 15	-0.01584	0.06435	-0.04739	0	0.02493	0.01128	0.02445	-0.05956	0	-0.02544	-0.01948	0	0.10989	-0.05737	1	0.05203	0.0894	0.03982	0	-0.02708	-0.0163	-0.03415	-0.02
Topic 16	0.04402	-0.0204	0	0.0491	0.01132	0.01657	-0.01901	0.05008	-0.02052	0.02085	0	0.08361	-0.02325	0	0.05203	1	0.01646	0.11518	-0.016	0	-0.03484	0	-0.01
Topic 17	-0.02831	0.23075	-0.0289	-0.02598	0.07845	-0.02008	0.07195	-0.04531	-0.03674	0	0	-0.01692	0.0888	-0.05324	0.0894	0.01646	1	-0.02444	-0.02215	-0.01591	0	-0.01942	-0.01
Topic 18	-0.01358	-0.02551	0	0.07077	0	0.12345	-0.02402	-0.02691	0.07217	-0.031	-0.02921	0.08006	-0.05918	0.02259	0.03982	0.11518	-0.02444	1	-0.04835	0	-0.03541	-0.02225	-0.04
Topic 19	-0.01484	-0.02324	0	-0.02306	0	-0.0336	-0.01592	-0.0143	0	-0.02694	-0.02911	-0.031	-0.02678	0.01696	0	-0.016	-0.02215	-0.04835	1	-0.03246	-0.02508	-0.03831	-0.02
Topic 20	0	-0.02835	0	0.02752	-0.01243	0.162	-0.0134	-0.01575	-0.02592	0	0.06049	0.02484	-0.04556	0	-0.02708	0	-0.01591	0	-0.03246	1	-0.0252	-0.01298	-0.03
Topic 21	-0.03806	-0.01444	-0.02887	-0.02562	-0.05015	-0.03797	0	-0.04765	0	-0.02591	-0.02156	-0.02003	0.05906	-0.05609	-0.0163	-0.03484	0	-0.03541	-0.02508	-0.0252	1	-0.0215	0.05
Topic 22	0.0566	0	-0.01814	0	-0.01833	-0.03375	0	0.13539	-0.03426	0.0907	0	0	0	-0.03607	-0.03415	0	-0.01942	-0.02225	-0.03831	-0.01298	-0.0215	1	-0.02
Topic 23	-0.03502	-0.02626	-0.03954	-0.024	-0.0766	-0.01242	-0.01915	-0.03298	0	-0.04165	0	0.02765	0.09609	-0.09747	-0.02979	-0.01056	-0.01467	-0.04621	-0.02316	-0.03005	0.05869	-0.02486	
Topic 24	0.05022	-0.0421	0.04072	0.09958	-0.05892	0.11461	-0.02217	0.01372	0	-0.02213	-0.02587	0.02424	-0.03783	-0.01531	0.03247	0.07406	-0.02031	0.09412	-0.04009	0	-0.03348	-0.01742	0.01
Topic 25	-0.03888	0.1813	-0.03332	-0.03063	0.07947	-0.03822	0.10948	-0.0461	-0.02007	-0.02562	-0.02254	-0.02797	0.13732	-0.05495	0.104	-0.03361	0.20317	-0.04472	0	-0.03063	0	-0.02961	-0.01
Topic 26	-0.03385	0.04821	-0.02462	-0.02266	-0.01813	-0.03613	0.07348	-0.04115	-0.0276	-0.01858	-0.01763	-0.01843	0.1603	-0.04822	0.03371	-0.03649	0.04251	-0.03855	0	-0.02236	0.01004	-0.01766	
Topic 27	-0.04444	0.0373	-0.04431	-0.03205	-0.02023	-0.04915	0	-0.05999	0.01226	-0.03754	-0.03063	0	0.14771	-0.09425	0.0751	0.01246	0.05616	-0.01777	0.02795	-0.04384	0.05128	-0.03084	0.09
Topic 28	-0.03373	0.03017	-0.02512	-0.02205	0.01715	-0.02522	0	-0.03924	0	-0.01993	-0.01508	-0.01475	0.03481	-0.0423	0.03434	-0.02126	0.07104	-0.02512	-0.01192	-0.01895	0.01404	-0.02228	0.02
Topic 29	-0.01839	-0.01935	-0.03297	-0.01619	0.02575	-0.02756	-0.02054	-0.02046	0	-0.01984	0	0.05616	0	-0.0664	0.13774	0.09921	0.01347	0.06276	-0.03256	-0.03213	-0.02131	-0.01308	0.02
Topic 30	0	0	0	0	0	0.04155	0	-0.01676	0	-0.01237	0	0	0	0	0.01381	0	0	0	0.04539	0	0	-0.01411	
Topic 31	-0.02524	0.01223	-0.02068	-0.019	-0.0342	-0.03025	0.02672	-0.03581	-0.02145	-0.01795	-0.017	-0.01885	0.20004	-0.04225	0.03579	-0.02813	0.03513	-0.03145	-0.01744	-0.01917	0.01917	-0.01318	0.02
Topic 32	0.02324	-0.01437	0.02587	0	0.15563	0	-0.01635	0.04394	-0.03469	-0.01551	-0.02023	-0.01947	-0.05839	0.0565	-0.03644	0	0	-0.02812	0.03214	0	-0.04182	-0.01708	-0.07
Topic 33	-0.04423	0.02696	-0.04886	-0.04092	0.01282	-0.05482	-0.02907	-0.05719	0	-0.0556	-0.04794	-0.01056	-0.0742	-0.03498	-0.09185	-0.03514	-0.04283	0	-0.03142	-0.04585	-0.01921	-0.03881	-0.03
Topic 34	-0.01011	0	-0.01522	0	0.02374	0.01088	0	-0.02965	-0.02003	-0.01536	-0.01102	0	-0.02622	-0.02154	0.06179	-0.01436	0	0.01758	-0.02259	0	-0.01883	-0.01945	-0.03
Topic 35	-0.0605	-0.02909	-0.0425	-0.03411	-0.04985	-0.05303	-0.01739	-0.06508	0.13261	-0.04313	-0.03443	0	0.0356	-0.08299	-0.01784	-0.03358	-0.03301	0	-0.03312	-0.04021	0.07027	-0.03921	0.11
Topic 36	0	-0.0397	0.09678	0.04538	0.018	-0.0277	-0.02359	0.18017	-0.01952	0.07147	0	0.02137	-0.07699	0.08774	-0.04639	0.06799	-0.03498	0	-0.02135	0	-0.05048	0.05444	-0.06
Topic 37	0	-0.01414	-0.02049	0	0.01473	-0.02742	-0.01249	0	-0.01189	0	0	-0.01511	-0.03437	-0.03893	0.0141	0	-0.01491	-0.01944	-0.0241	0	-0.02177	-0.01082	-0.01

stm::topicCorr

Aggregating topics: topic correlations

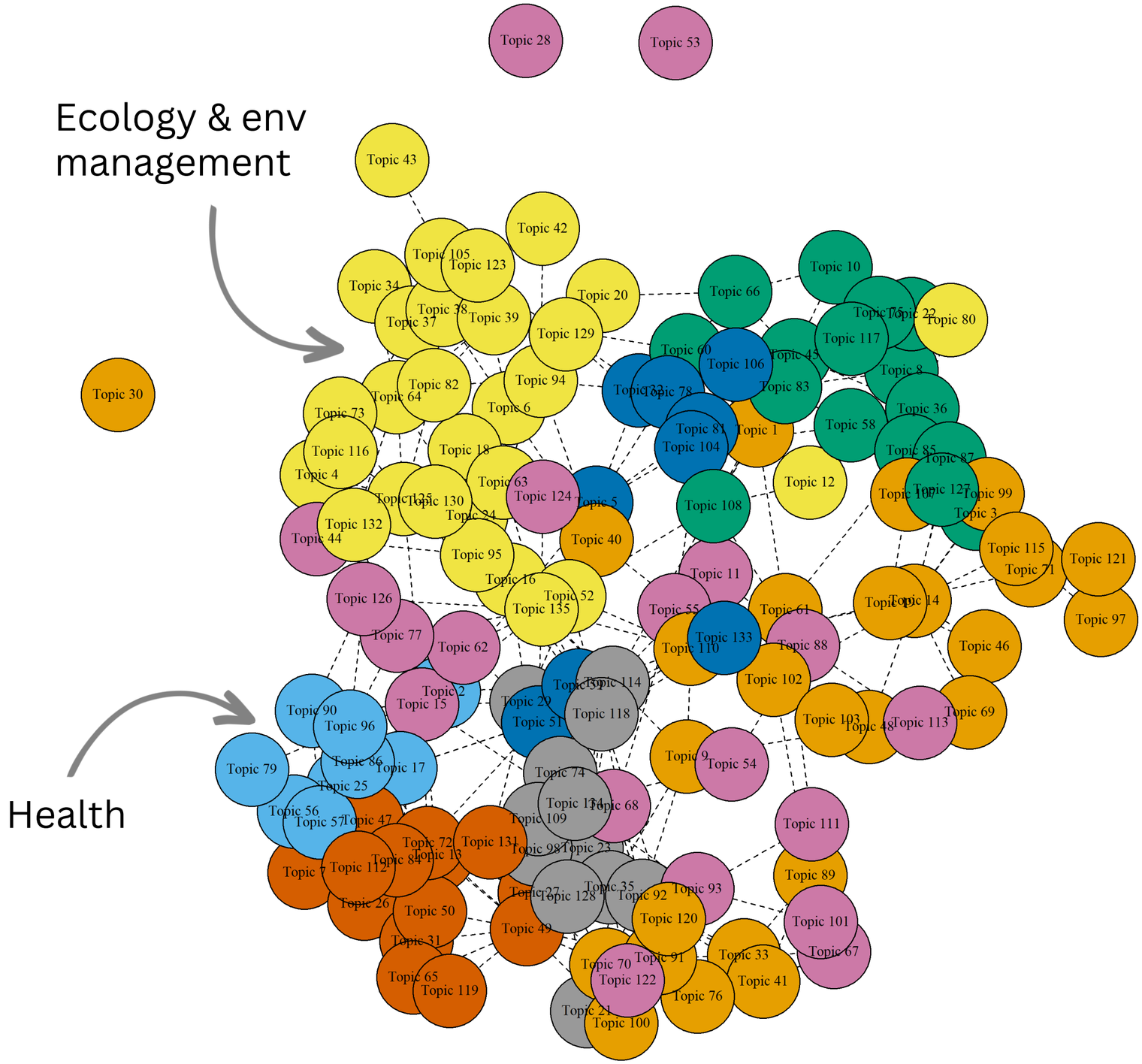


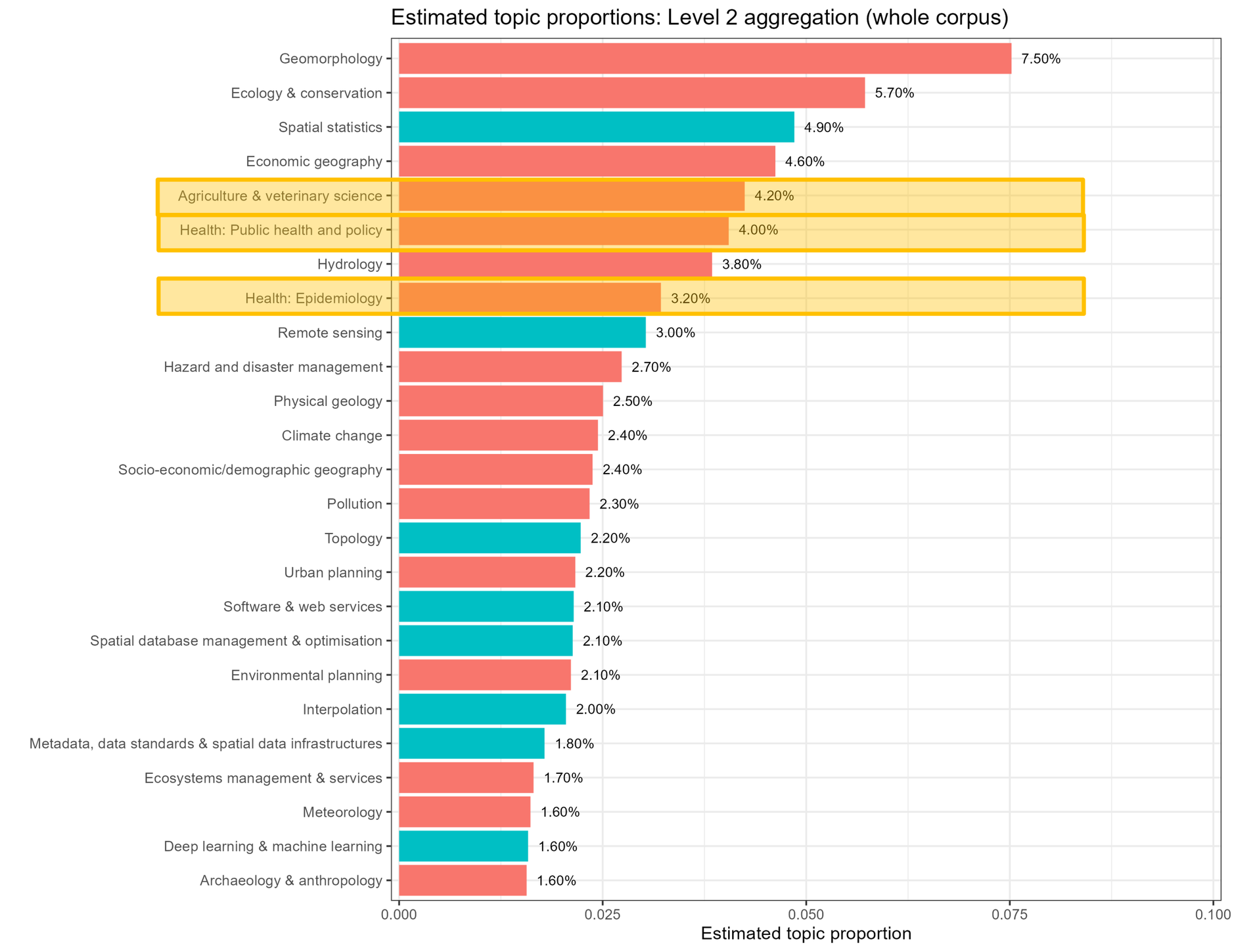
stm::topicCorr

factoExtra::fviz_dend

Aggregating topics: topic correlations

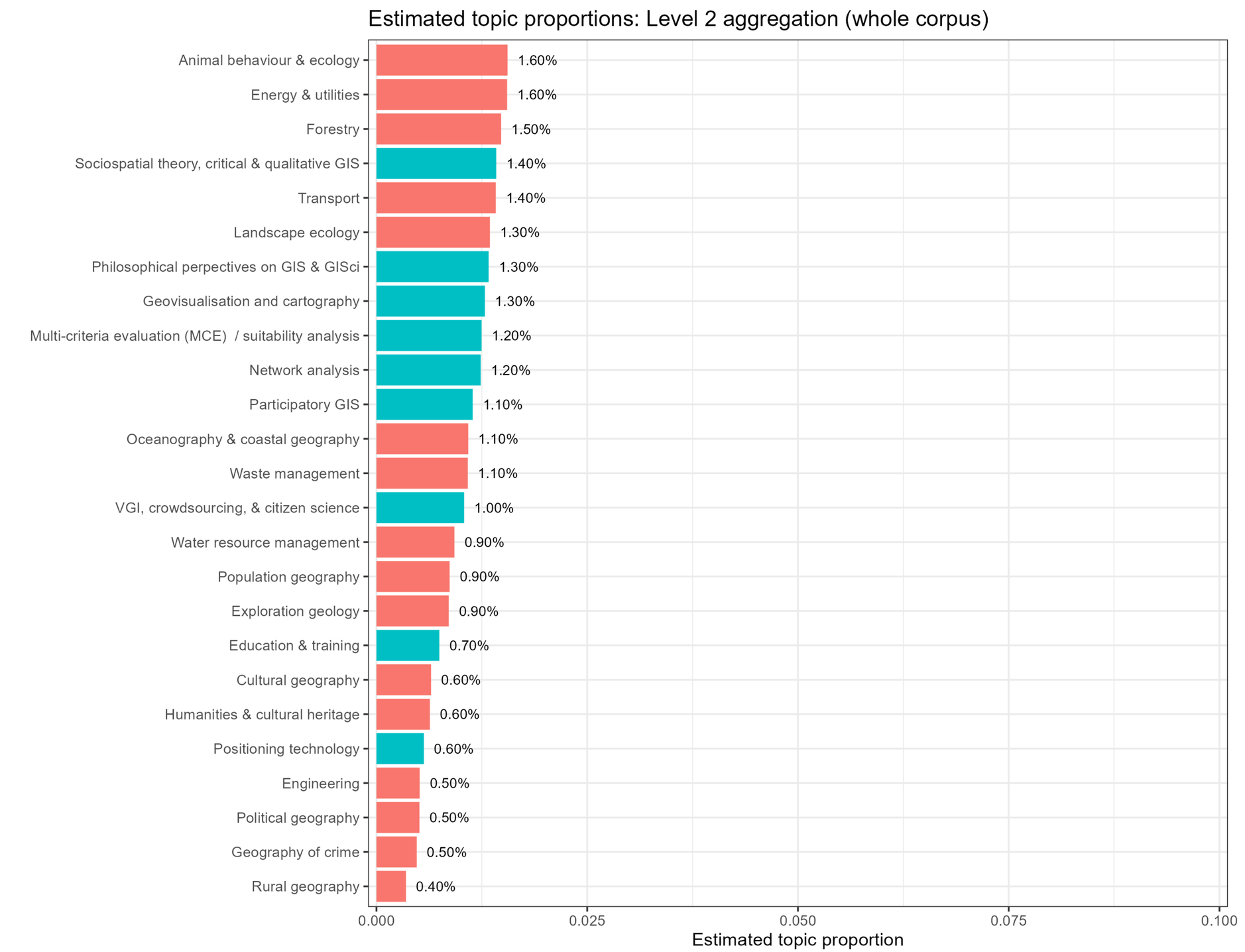
stm::topicCorr





Whole corpus
(all years)

Aggregated Topics
Level 2
(50 topics)

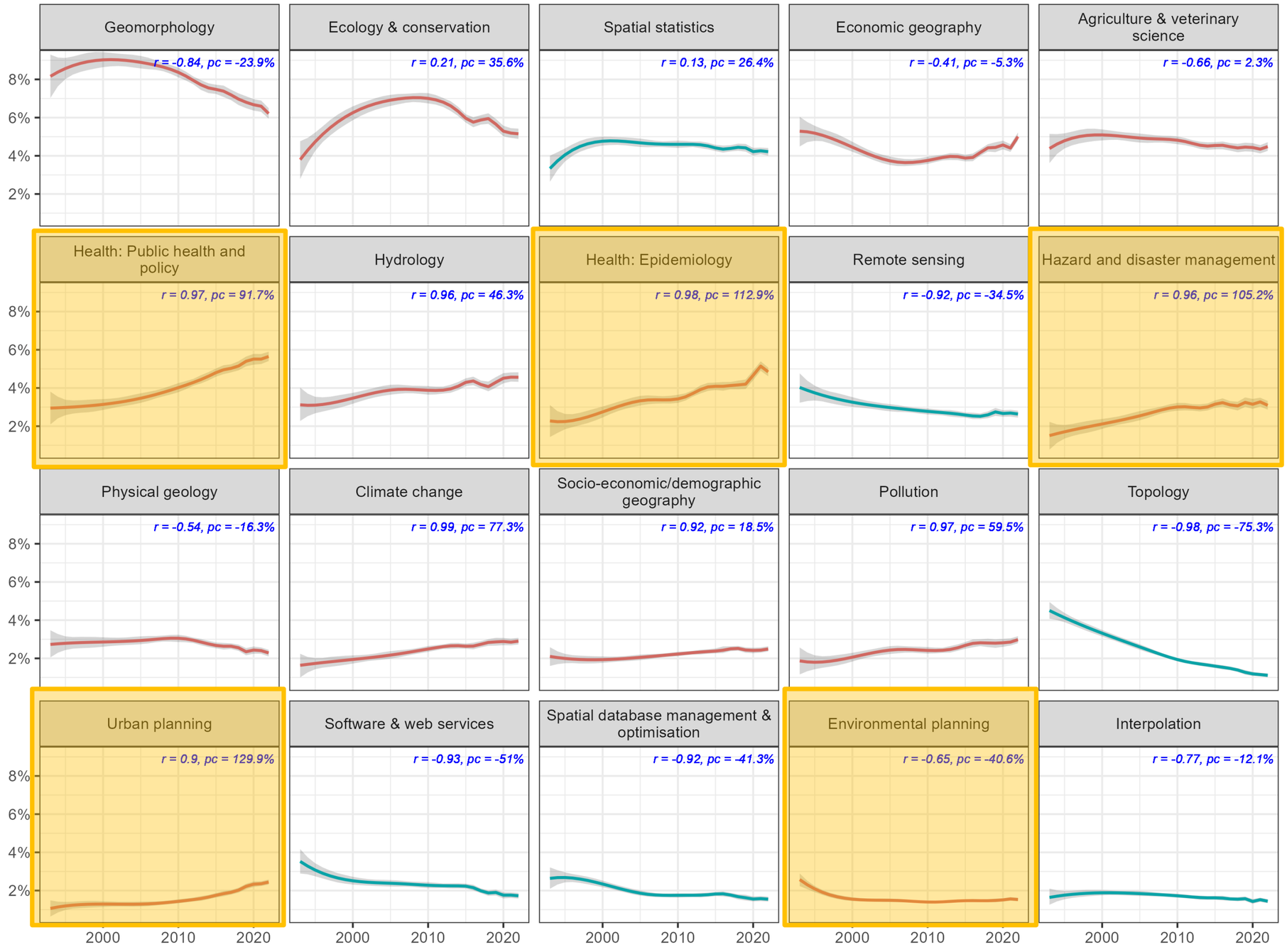


Whole corpus (all years)

Domain application

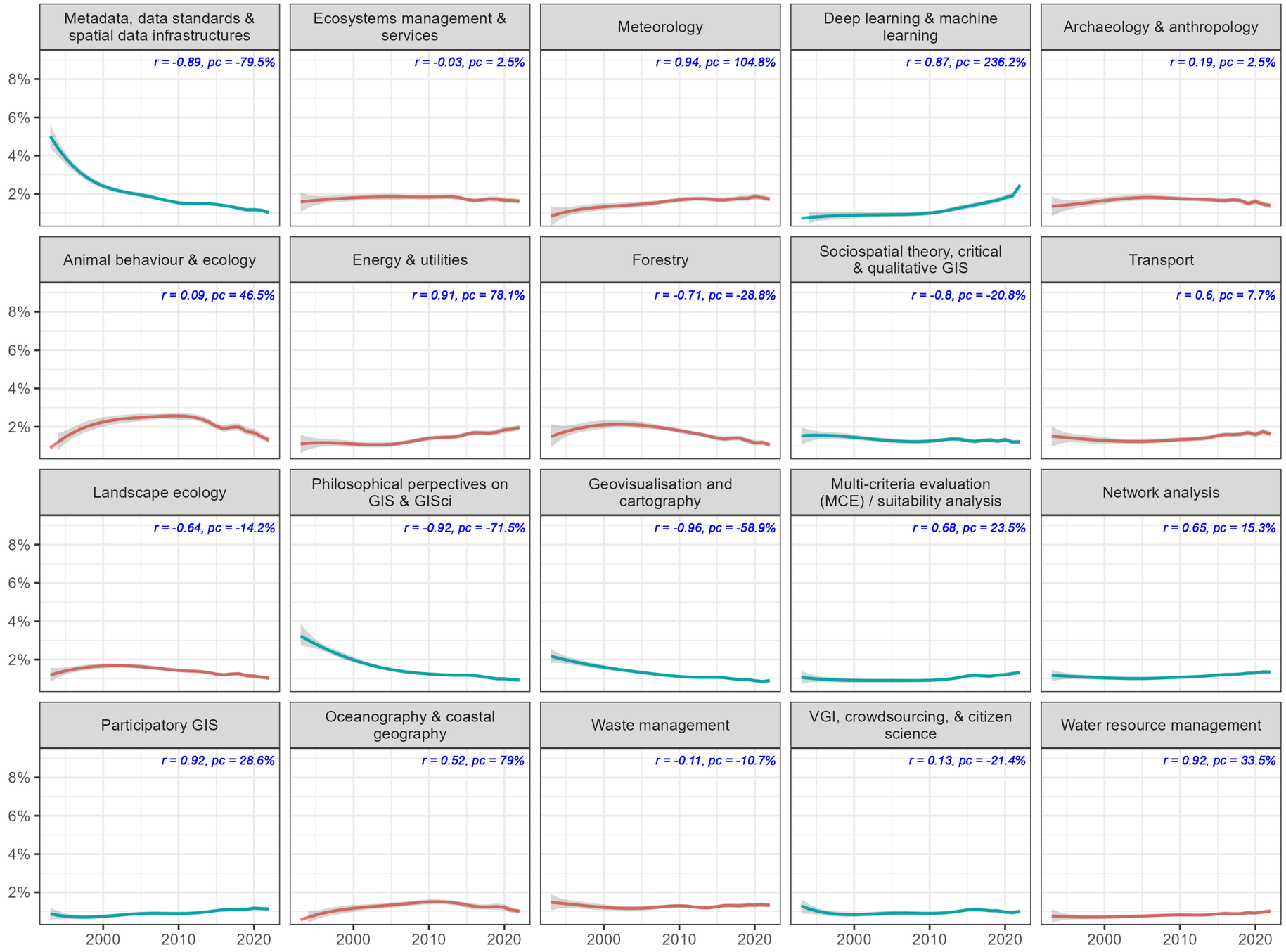
GISci

Aggregated Topics **Level 2**



Topics over time

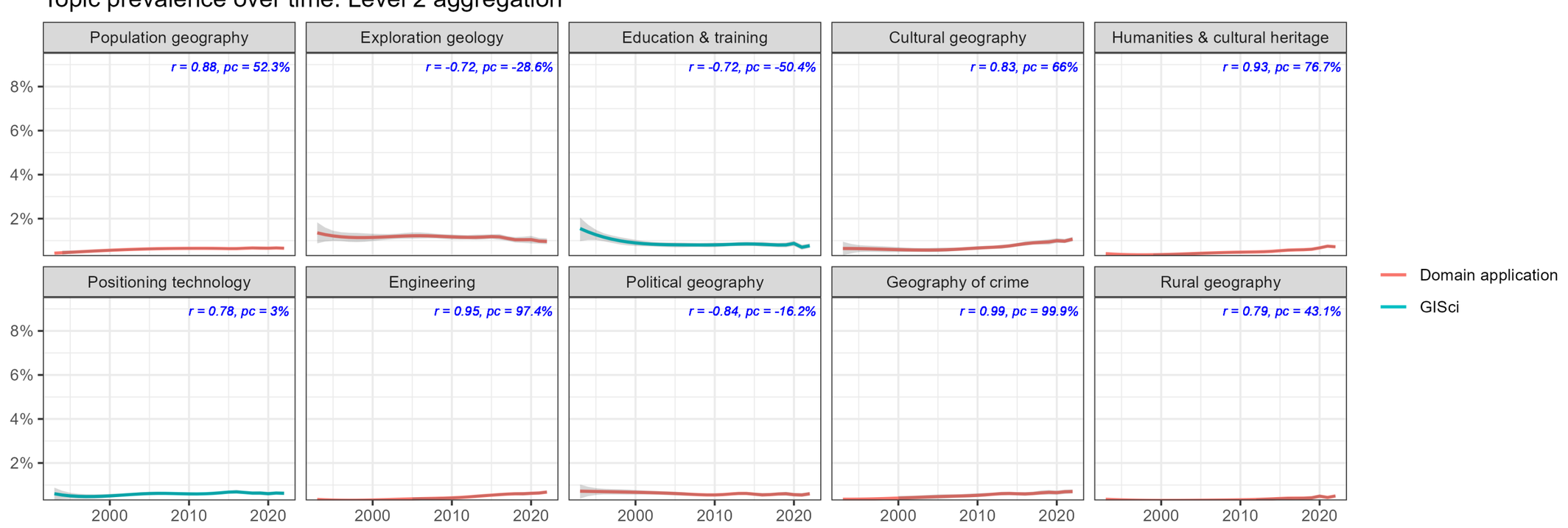
Aggregated Topics **Level 2**



Topics over time

Aggregated Topics

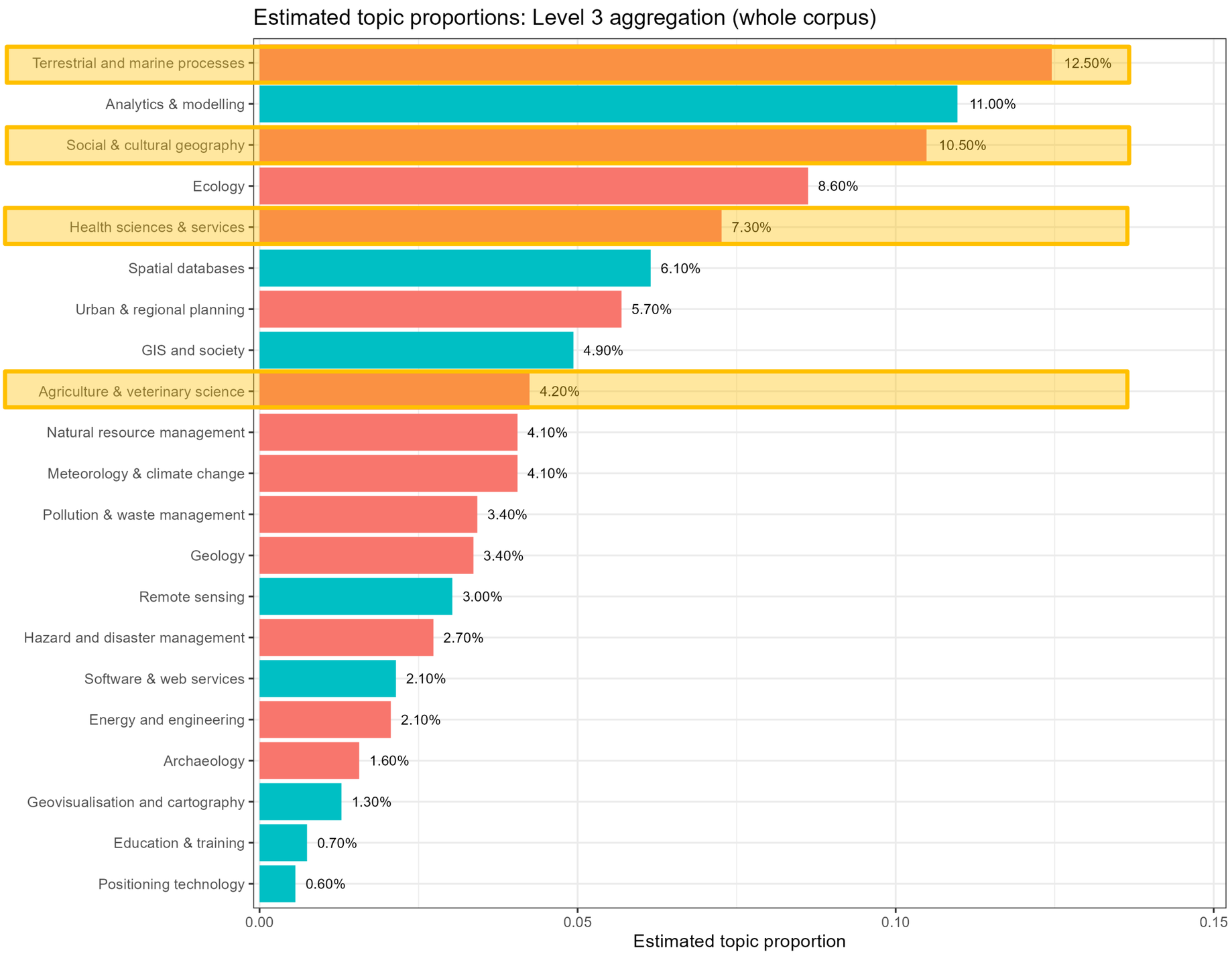
Level 2



Topics over time

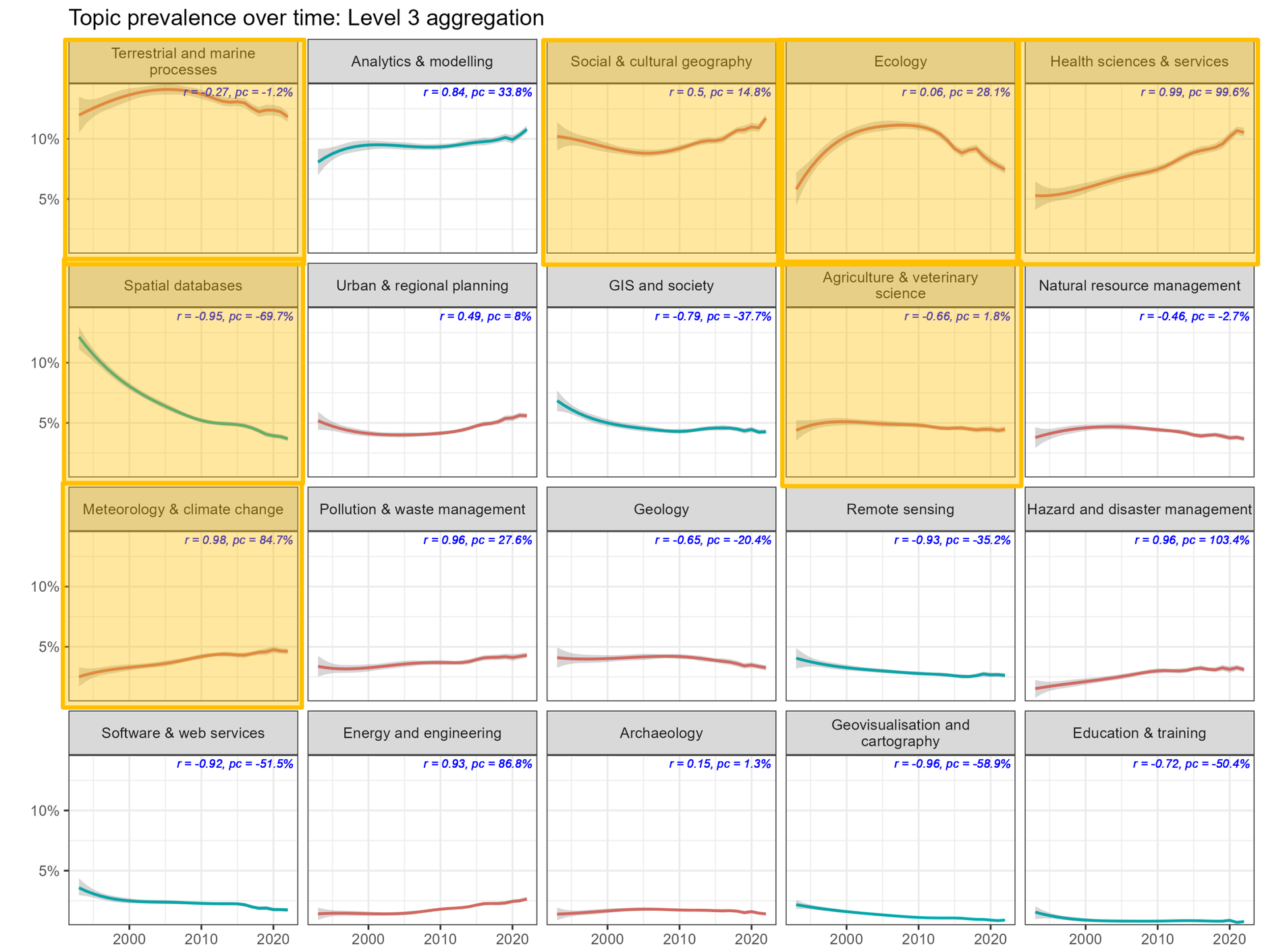
Aggregated Topics

Level 2



Whole corpus
(all years)

Aggregated Topics
Level 3
(21 topics)

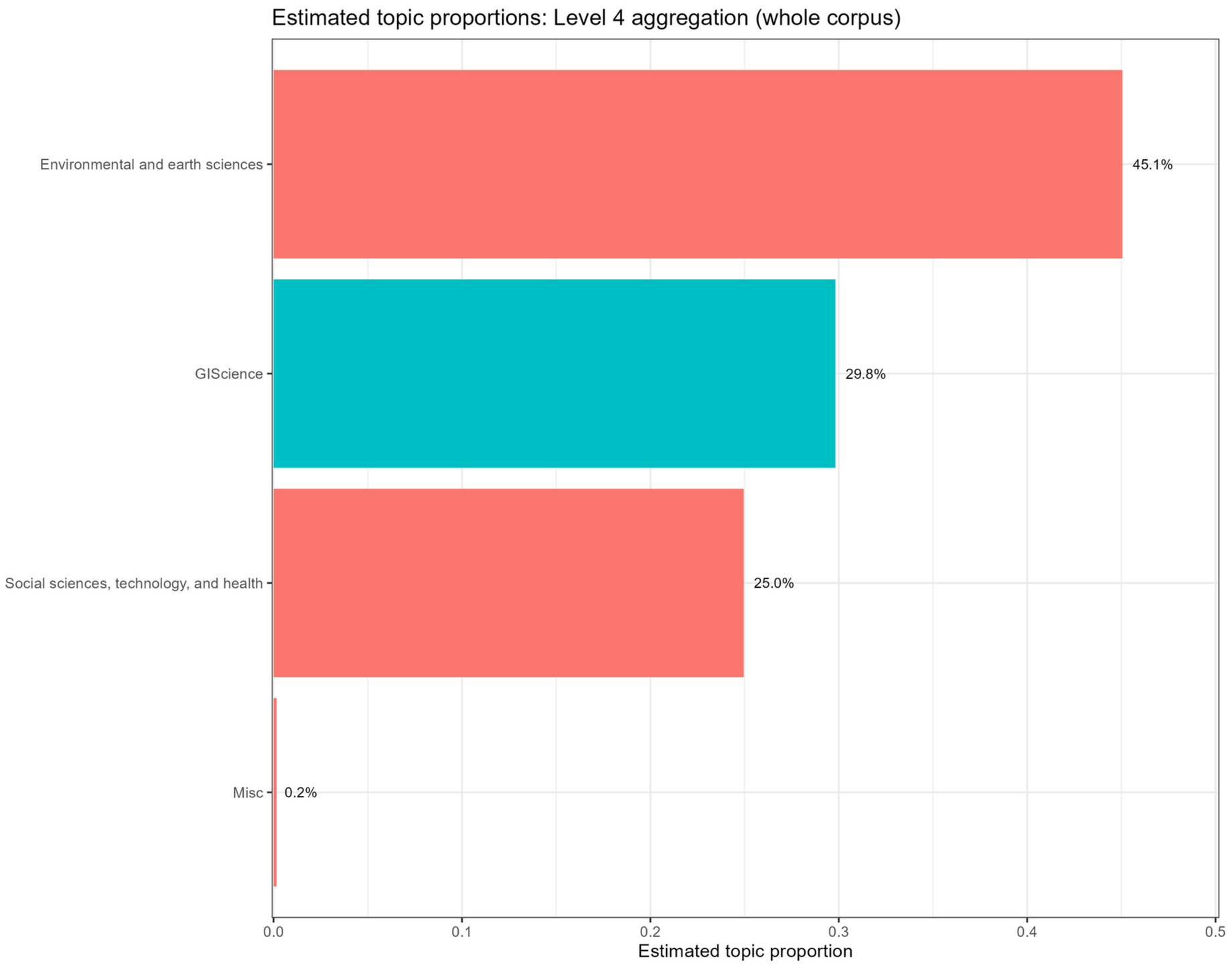


Topics over time

Domain application

GISci

Aggregated Topics **Level 3**

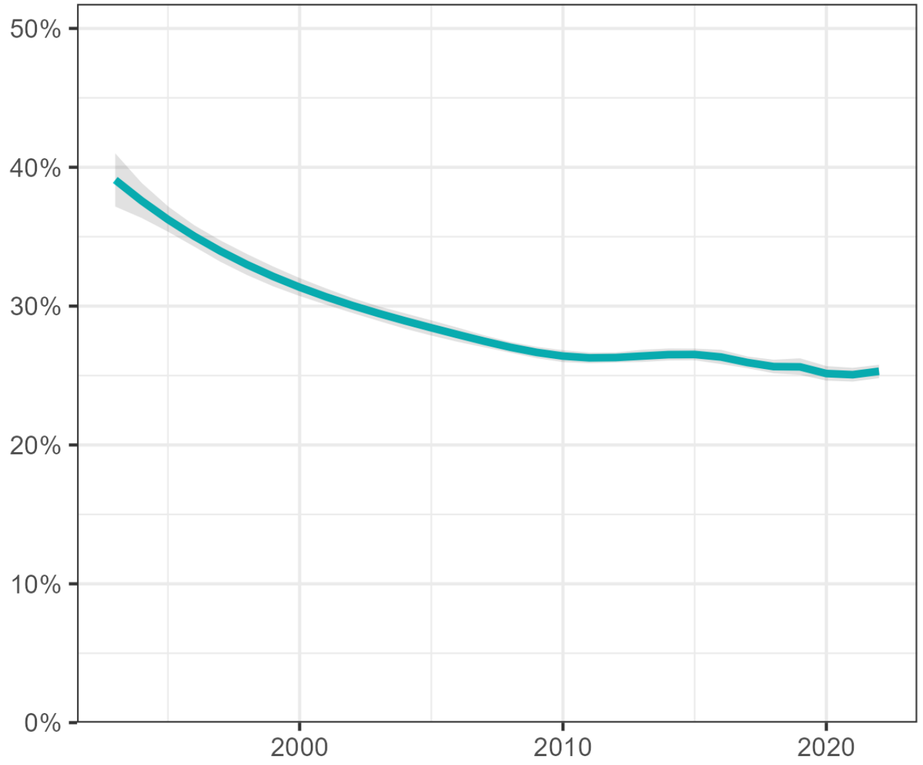


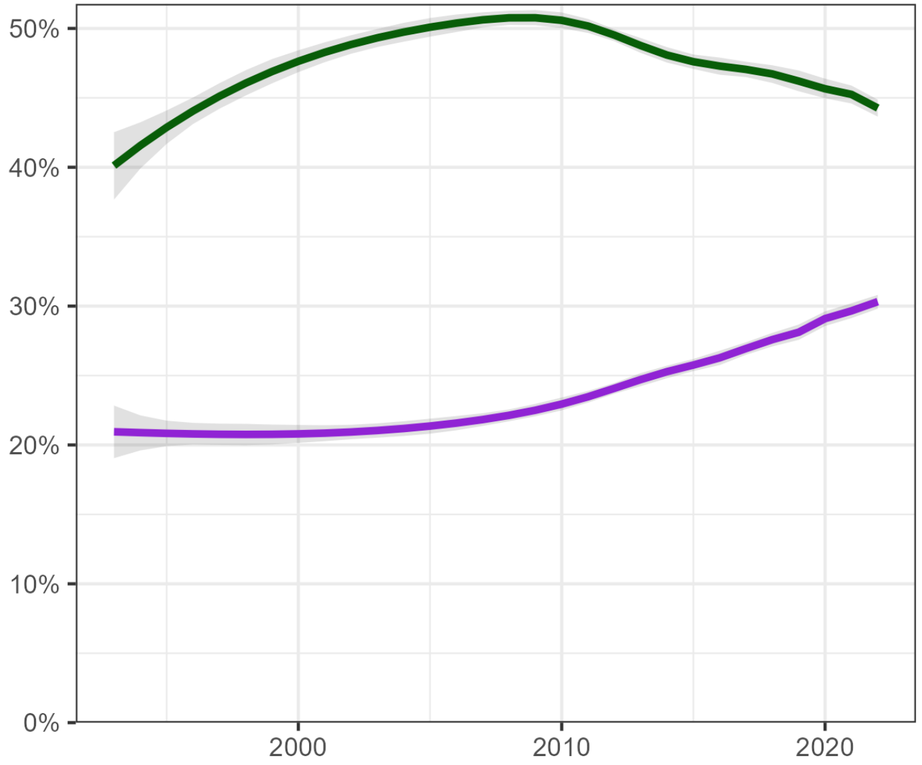
Whole corpus
(all years)

Aggregated
Topics
Level 4
(3 topics)

Aggregated Topics: Level 4

(Highest level of aggregation)





GIScience

Domain applications

Environmental and earth sciences

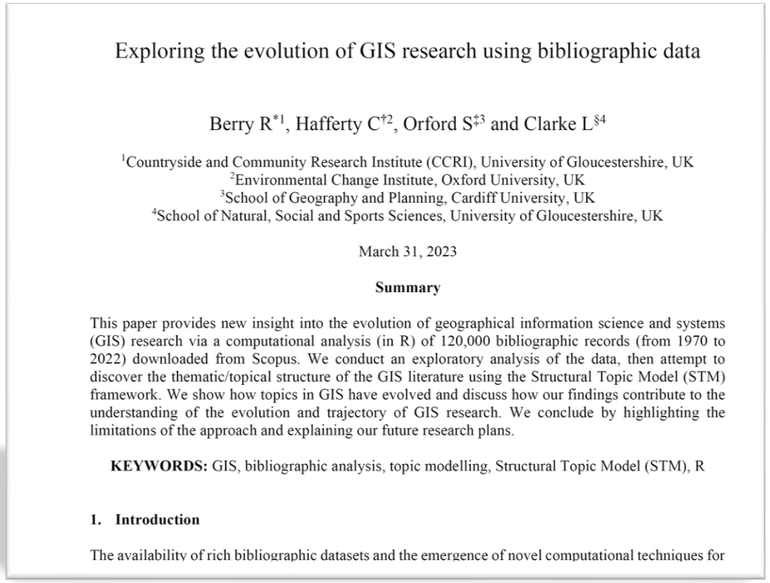
Social sciences, technology, and health

Summary

- Most comprehensive computational thematic analysis of GIS (any?) lit
- Provides a wealth of new insight into the evolution of GISR
  - GISci / GIS (applications)
  - Evolution and major trends/turning points
  - Responding to societal challenges (health, climate change, urbanization, economic inequality)
- **Limitations**
  - Single data source, biases English language
  - “Human in the loop” - no ‘best’ model
  - Aggregation difficult and imperfect - contentious
  - Computation improvements
  - Only one interpretation of the results
- **Open code** – allows other interpretations

Outputs

- **Conference paper**
  - GISRUK Conference, Glasgow, April 2023



Exploring the evolution of GIS research using bibliographic data

Berry R[*1], Hafferty C[†2], Orford S[‡3] and Clarke L[§4]

[1]Countryside and Community Research Institute (CCRI), University of Gloucestershire, UK
[2]Environmental Change Institute, Oxford University, UK
[3]School of Geography and Planning, Cardiff University, UK
[4]School of Natural, Social and Sports Sciences, University of Gloucestershire, UK

March 31, 2023

Summary

This paper provides new insight into the evolution of geographical information science and systems (GIS) research via a computational analysis (in R) of 120,000 bibliographic records (from 1970 to 2022) downloaded from Scopus. We conduct an exploratory analysis of the data, then attempt to discover the thematic/topical structure of the GIS literature using the Structural Topic Model (STM) framework. We show how topics in GIS have evolved and discuss how our findings contribute to the understanding of the evolution and trajectory of GIS research. We conclude by highlighting the limitations of the approach and explaining our future research plans.

KEYWORDS: GIS, bibliographic analysis, topic modelling, Structural Topic Model (STM), R

1. Introduction

The availability of rich bibliographic datasets and the emergence of novel computational techniques for

- **Journal papers**
  1. Cartography and Geographic Information Science – This work (thematic analysis)
  2. Journal of Maps – Geographical analysis
  3. International Journal of Geographical Information Science – Spatio-thematic/impact analysis
- **Wider CCRI research**

Thank you!

rberry@glos.ac.uk

@rural_gis

CCRI
COUNTRYSIDE AND COMMUNITY RESEARCH INSTITUTE

UNIVERSITY OF OXFORD

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

UNIVERSITY OF GLOUCESTERSHIRE