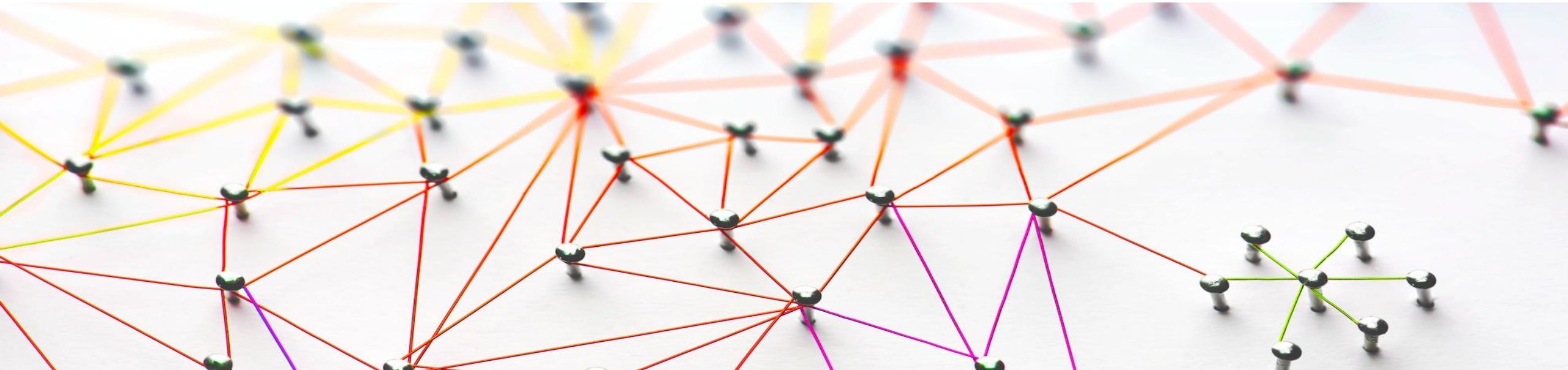
UFZ HELMHOLTZ Centre for Environmental Research

Automatized drought impact detection from newspaper articles using natural language processing and machine learning

Jan Sodoge, Dr. Mariana Madruga de Brito, Prof. Christian Kuhlicke

Department of Urban and Environmental Sociology



Manifold drought impacts and their assessment

Drought impacts observed in Germany 2018/2019

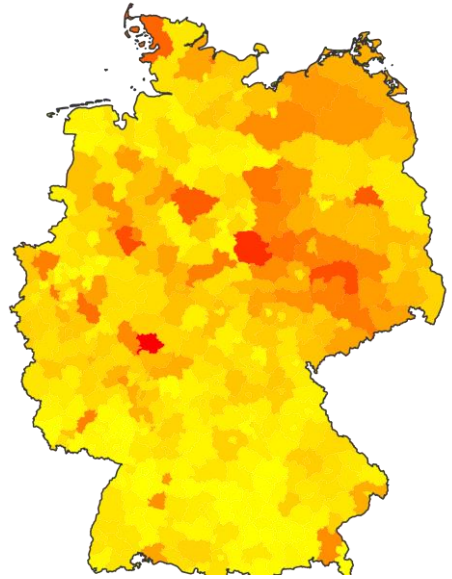
de Brito, M. M., Kuhlicke, C., & Marx, A. (2020). Near-real-time drought impact assessment: a text mining approach on the 2018/19 drought in Germany. *Environmental Research Letters*, *15*(10), 1040a9.

Leveraging machine learning and natural language processing to mine drought impacts

Sample collection

Machine learning

Natural language processing

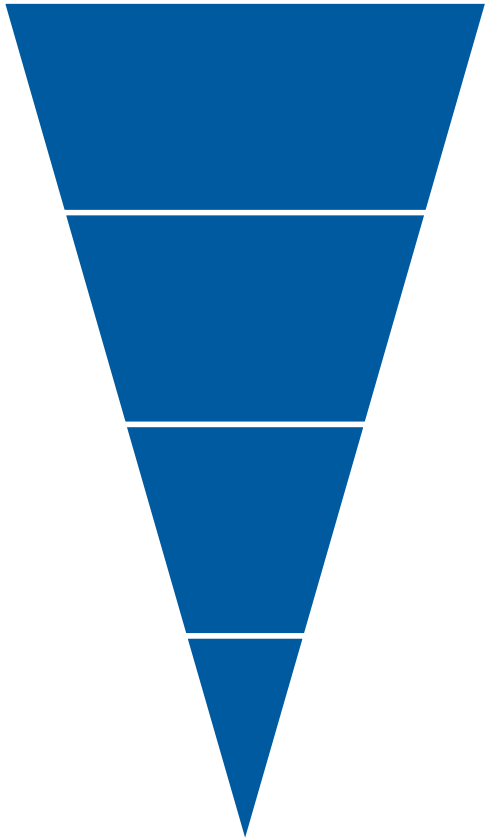


Newspaper articles covering Germany between 2000-2021

Drought impact statements
- *What type of drought impact?*
- *Where was the impact?*
- *When did it happen?*

Spatial-temporal drought impact dataset

Sample collection → removing duplicate & non-relevant information



Download articles from news-aggregator database using keywords

Calculate Jaccard similarity between all articles, remove duplicate articles with similarity > 0.90

Detect articles with non-relevant topics using a topic modelling approach

Text corpus for drought impact analysis

Drought impact classification: Turning text to logistic regression

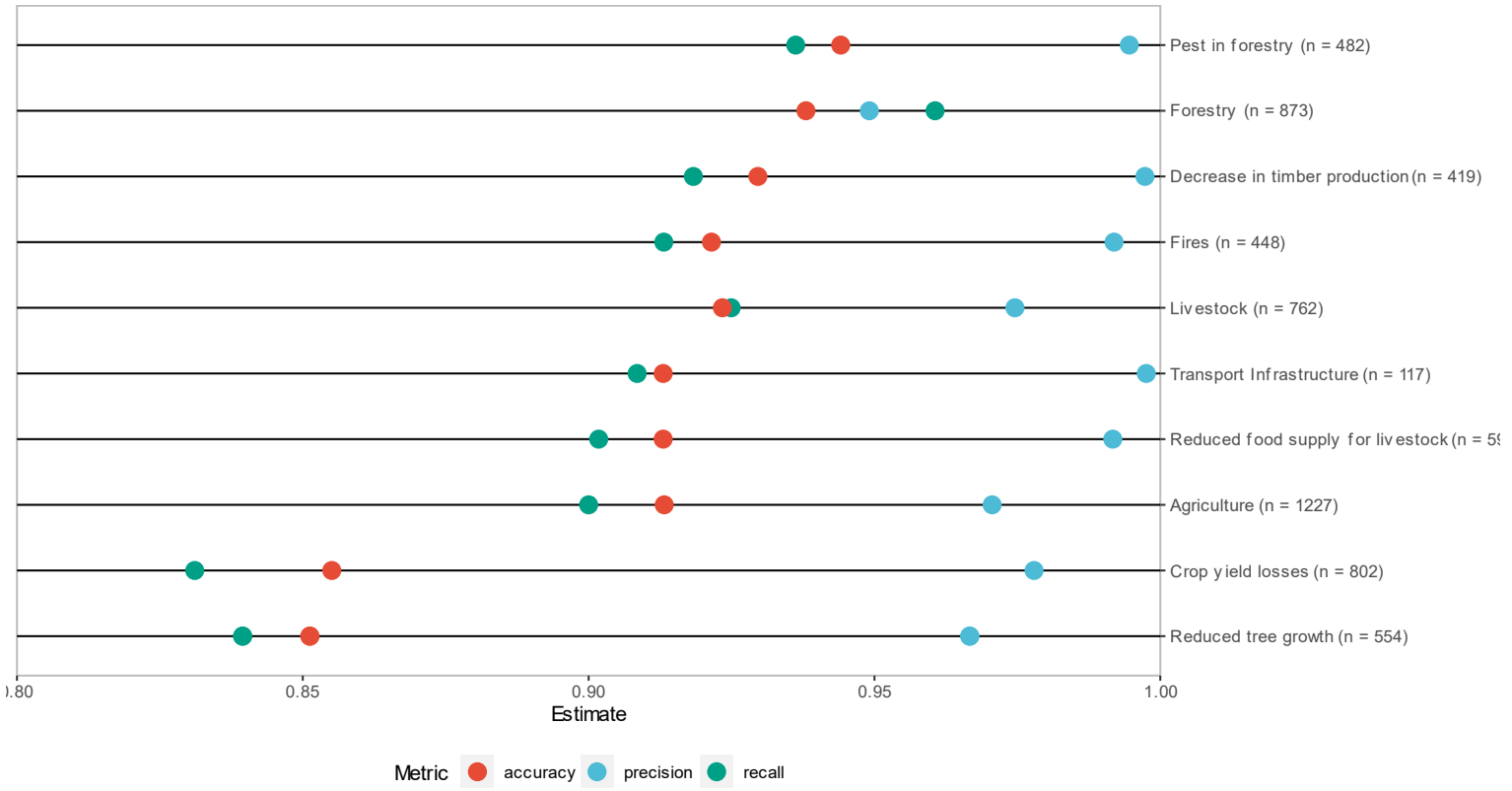
Text in classification models?

"Because of the drought the farmers cannot avoid reduced yields"

Words ('tokens') as variables

Lasso logistic regression

Impact observed ←→ no impact observed

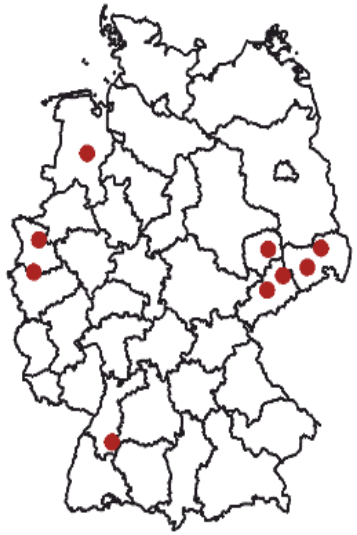


Location identification: Examining an article's geographic scope

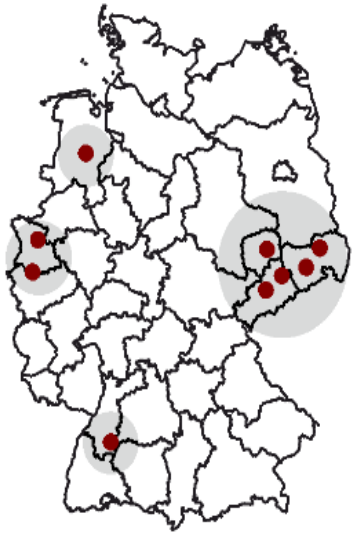
Detection of potential locations in text using

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea

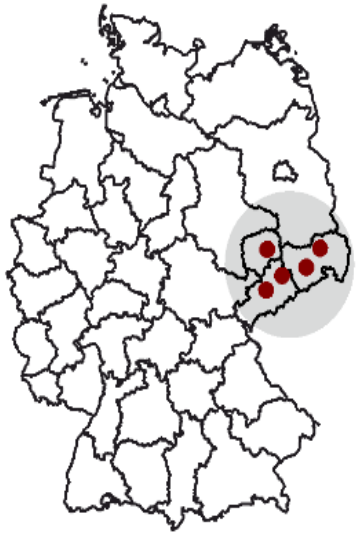
Geoparsing potential locations via gazetteer data



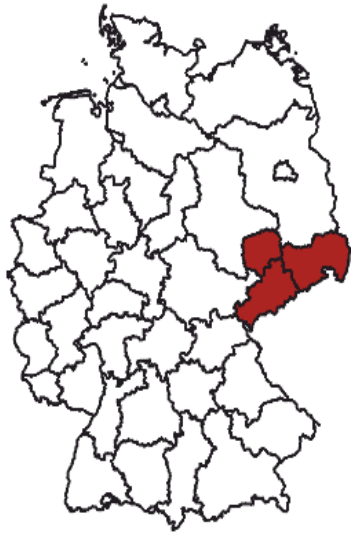
Sorting potential locations in clusters



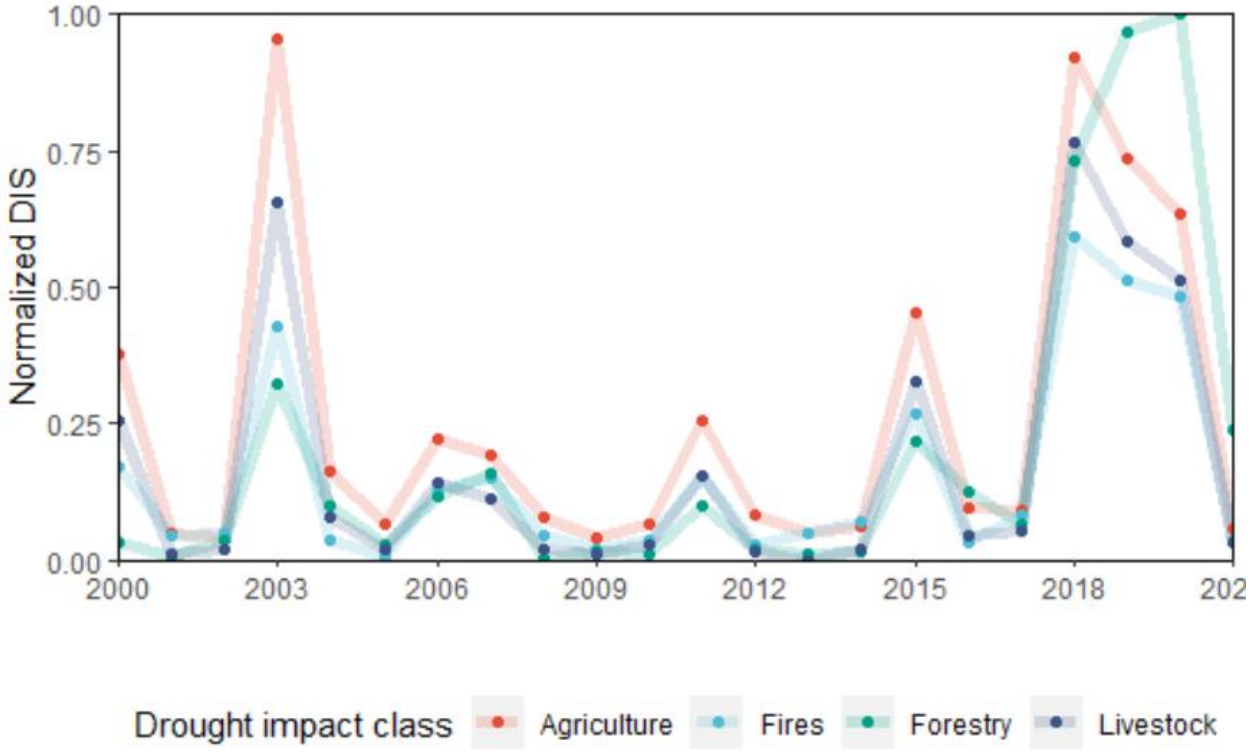
Selecting one cluster as the drought impact location area

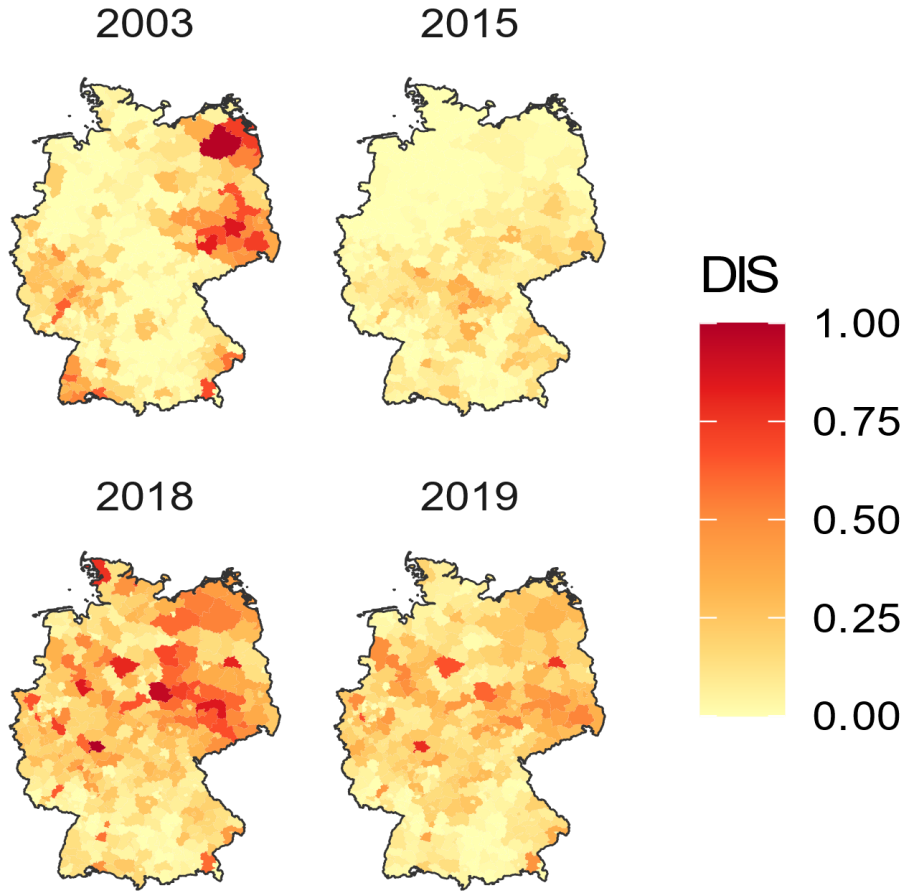


Mapping drought impact locations to statistical units



A spatial-temporal perspective on drought impacts in Germany





Empirical validation of the resulting dataset

- **Does the resulting dataset match the trends of external indicators?**
  - Validation with both bio-physical measures (precipitation deficit) and impact-specific indicators (e.g. forest fire statistics)
  - Correlation analysis concerning both spatial (in each year) and temporal trends (in each spatial unit)
- → **General spatial and temporal trends reflected across all indicators**
  - Years with drought events perform better on average

Conclusion and outlook

- Automatization of a yet manual task in drought impact assessment, while maintaining high levels of accuracy
- → remove limitations to the number of processed texts
  → larger datasets
- Easy adaptation to other domains, hazard types, or case studies (e.g. floods, training on other languages)
- Dataset potentials: ML to model linkages with bio-physical drought indicators

Thank you for having me

Questions, comments, ideas?
Or drop me a message

 jan.sodoge@ufz.de

 Twitter: @jsodoge