The potential of web scraping for border regional analyses: The example of real estate data

Elias Günther, Stefan Hippe, Tobias Chilla

Data availability in border regions remains a pressing issue – in particular with regard to harmonisation issues. In the absence of harmonized data, so called new data can be a helpful solution. Digitalisation comes along with a series of new data sources and allows for new approaches to study the dynamics of cross-border regions. Useful data can be collected via various methods, including web scraping and openly available data products supplied by data owners (e.g., travel information from private companies). Web scraping is a method that allows for automatic extraction of relevant information from online media (websites, application platforms, open databases). This is an important complement to traditional data collection methods (e.g., surveys, census).

In our paper, we explore the potentials and limitations of web scraping for border regional analyses with the example of real estate data. Through web scraping, absolute and relative price indicators and quality related indicators can be taken on board. Differentials and integration patterns will be revealed regarding rental and buying costs, both in relation to distance to the border. In addition, the number of offers gives an indication of the housing market dynamics. This data can be a proxy to understand dynamics and drivers of cross-border commuting, particularly in border regions with urban centres in close proximity to the border.

Data are taken from the research project on “The potential of new data sources for border-related knowledge gaps (BorderData)” funded by the Bavarian-Czech Research Agency (BTHA), focussing on the Bavaria/Czech region. The presentation reflects a) technical options and barriers in organising automated data retrieval, b) on data quantity and availability, c) on data quality and potential corrections and calibration options.