# From Rural Roots to Digital Fruits: AI-Driven Insights into Innovation

*Korneliusz Pylak,*
*Lublin University of Technology, POLAND*

**Keywords**

Rural areas
Innovation
Unstructured Website Text
Spatial Dynamics
Natural Language Processing (NLP)
Large Language Models (LLM)

**JEL codes**

O31 - Innovation and Invention: Processes and Incentives
O32 - Management of Technological Innovation and R&D
C55 - Large Data Sets: Modelling and Analysis
L86 - Information and Internet Services; Computer Software
R11 - Regional Economic Activity: Growth, Development, Environmental Issues, and Changes

## Extended abstract

Traditional methodologies for estimating innovation and firm performance have traditionally relied on established sources of secondary data, such as patents, academic publications, R&D projects and administrative records (Abbasiharofteh et al, 2022; Cillo et al, 2019; Nasirov, 2020; Simensen and Abbasiharofteh, 2022). However, the geography of innovation is undergoing a transformation, particularly in rural areas, where patents and research projects may be scarce or, conversely, innovation may be mostly localised (Fritsch and Wyrwich, 2021), posing a challenge in measuring this phenomenon. However, recent advances in computational power and language modelling, coupled with large amounts of textual data from job postings, patent documents, online texts and trademark records, are opening up new opportunities to

study regional economic development, labour market dynamics and the geography of knowledge production (Aweisi et al, 2021; Cetera et al, 2022; Gök et al, 2015; Skhvediani et al, 2022).

This digital revolution is particularly relevant for rural regions, where traditional innovation indicators are often inadequate due to sparse data. Our research makes a pioneering contribution by offering a novel approach to measuring rural innovation. This method not only deepens our understanding of rural innovation systems, but also equips policy makers and stakeholders with a powerful tool to design targeted interventions that support rural entrepreneurship, mitigate economic disparities and promote sustainable industrial transformation – critical factors for driving economic development in rural areas.

Our approach utilises the registration data of companies located in rural areas, identifies their websites and applies advanced unstructured textual data analysis to extract information about their products and services. Going beyond static snapshots, our methodology includes not only techniques such as web scraping, social network analysis and natural language processing (Abbasiharofteh et al, 2023; Ashouri et al, 2022; Kinne & Axenbeck, 2020; Skhvediani et al, 2022). Indeed, we also introduce a temporal dimension – tracking changes in the content of web pages archived by the Wayback Machine. We are thus able to detect the introduction of new products and services, thus signalling product innovation at the company level. While previous research has used textual predictions in contexts such as analyst reports (Bellstam et al., 2021), our innovative use of unstructured website text allows us to track the evolution of innovative activities in a more comprehensive way.

Furthermore, we recognise that a firm's ability to innovate is closely related to its ability to integrate existing knowledge and resources over time (Audretsch & Belitski, 2022; Bruno et al., 2022; Tomizawa et al., 2020). The spatial dimension is particularly important here, as physical proximity and relationships between firms facilitate knowledge exchange and catalyse innovation (Alam et al., 2022; Bailey et al., 2018; Obschonka et al., 2023; Singh et al., 2022). Given the co-location of innovative actors, our approach captures dynamic spatial interactions and maps digital footprints to reveal how local knowledge diffusion and rural innovation systems evolve. This sophisticated understanding of the geography of knowledge production allows us to identify product innovations at both local and regional scales – and, with further scaling, at the national level and beyond.

We test our methodology in Poland, starting with a comprehensive REGON database that catalogs all Polish companies, including farms, sole proprietors and legal entities. In order to focus on innovation in rural areas, we exclude farms unless their owners are engaged in additional economic activity (identified by a common tax identification number) and remove companies located in Poland's 930 cities based on geolocation data provided in the registry.

For each selected rural company, we identify and retrieve its websites – either from records or through targeted web searches. We then implement advanced web scraping techniques to extract unstructured textual data that describes the company's products and services in detail. Using a state-of-the-art Large Language Model (LLM), we automatically categorize and extract product names and descriptions, thereby creating a comprehensive database that links companies to their websites, product titles, descriptions and the specific websites where this information is found. We then use the Wayback Machine to determine the earliest appearance of these website links, effectively timing the introduction of new products and services.

In the second stage, we fine-tune our LLM to recognize and prioritize links containing information about products and services. By integrating advanced techniques—such as transformer-based semantic analysis, named entity recognition and embeding-based clustering—we speed up the search process in Wayback Machine, significantly reducing the time required to extract historical content. As a result, we only extract content of historical pages classified as product/service pages. This approach enables us to compile a full timeline of product and service launches, capturing the dynamics of innovation both locally and regionally.

Moreover, our methodology sets the stage for future extensions. By scaling these techniques, researchers can study urban innovation ecosystems and analyze the diffusion of innovation between cities and rural areas. Overall, our approach not only maps the trajectory of innovation in rural regions with unprecedented detail, but also provides policymakers and stakeholders with a powerful framework to drive sustainable industrial transformation.

## Bibliography


Abbasiharofteh, M., Castaldi, C., & Petralia, S. (2022). *From patents to trademarks: Towards a concordance map*. European Patent Office.

Abbasiharofteh, M., Krüger, M., Kinne, J., Lenz, D., & Resch, B. (2023). The digital layer: Alternative data for regional and innovation studies. *Spatial Economic Analysis*, 1–23.

Alam, M. A., Rooney, D., & Taylor, M. (2022). From ego-systems to open innovation ecosystems: A process model of inter-firm openness. *Journal of Product Innovation Management*, *39*(2), 177–201.

Ashouri, S., Suominen, A., Hajikhani, A., Pukelis, L., Schubert, T., Türkeli, S., Van Beers, C., & Cunningham, S. (2022). Indicators on firm level innovation activities from web scraped data. *Data in Brief*, *42*, 108246.

Audretsch, D. B., & Belitski, M. (2022). The knowledge spillover of innovation. *Industrial and Corporate Change*, *31*(6), 1329–1357.

Aweisi, A., Arora, D., Emby, R., Rehman, M., Tanev, G., & Tanev, S. (2021). Using web text analytics to categorize the business focus of innovative digital health companies. *Technology Innovation Management Review*, *11*(7/8).

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives*, *32*(3), 259–280.

Bellstam, G., Bhagat, S., & Cookson, J. A. (2021). A Text-Based Analysis of Corporate Innovation. *Management Science*, *67*(7), 4004–4031.

Bruno, R. L., Crescenzi, R., Estrin, S., & Petralia, S. (2022). Multinationals, innovation, and institutional context: IPR protection and distance effects. *Journal of International Business Studies*, *53*(9), 1945–1970.

Cetera, W., Gogołek, W., Żołnierski, A., & Jaruga, D. (2022). Potential for the use of large unstructured data resources by public innovation support institutions. *Journal of Big Data*, *9*(1), 46.

Cillo, V., Petruzzelli, A. M., Ardito, L., & Del Giudice, M. (2019). Understanding sustainable innovation: A systematic literature review. *Corporate Social Responsibility and Environmental Management*, *26*(5), 1012–1025.

Daas, P. J., & van der Doef, S. (2020). Detecting innovative companies via their website. *Statistical Journal of the IAOS*, *36*(4), 1239–1251

Fritsch, M., & Wyrwich, M. (2021). Does Successful Innovation Require Large Urban Areas? Germany as a Counterexample. *Economic Geography*, 97(3), 284–308. https://doi.org/10.1080/00130095.2021.1920391.

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671.

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, *125*(3), 2011–2041.

Nasirov, S. (2020). Trademark value indicators: Evidence from the trademark protection lifecycle in the U.S. pharmaceutical industry. *Research Policy*, *49*(4), 103929.

Obschonka, M., Tavassoli, S., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2023). Innovation and inter-city knowledge spillovers: Social, geographical, and technological connectedness and psychological openness. *Research Policy*, *52*(8), 104849.

Simensen, E. O., & Abbasiharofteh, M. (2022). Sectoral patterns of collaborative tie formation: Investigating geographic, cognitive, and technological dimensions. *Industrial and Corporate Change*, *31*(5), 1223–1258.

Singh, A., Chhetri, P., & Padhye, R. (2022). Modelling inter-firm competitive rivalry in a port logistics cluster: A case study of Melbourne, Australia. *The International Journal of Logistics Management*, *33*(2), 455–476.

Skhvediani, A., Sosnovskikh, S., Rudskaia, I., & Kudryavtseva, T. (2022). Identification and comparative analysis of the skills structure of the data analyst profession in Russia. *Journal of Education for Business*, *97*(5), 295–304.

Tomizawa, A., Zhao, L., Bassellier, G., & Ahlstrom, D. (2020). Economic growth, innovation, institutions, and the Great Enrichment. *Asia Pacific Journal of Management*, *37*(1), 7–31.