ISSN 1611-1699 / eISSN 2029-4433

VILNIUS TECH
Vilnius Gediminas Technical University

JOURNAL of BUSINESS ECONOMICS & MANAGEMENT

2026
Volume 27
Issue 1
Pages 162–185

https://doi.org/10.3846/jbem.2026.25753

THE IMPACT OF FIRM-LEVEL INNOVATIVENESS ON SOCIOECONOMIC INDICATORS: AN AI-DRIVEN APPROACH

Alina STUNDŽIENĖ ✉, Vaida PILINKIENĖ, Mantas LUKAUSKAS, Andrius GRYBAUSKAS, Mantas VILKAS

School of Economics and Business, Kaunas University of Technology, Kaunas, Lithuania

Article History:
- received 25 August 2025
- accepted 19 December 2025

Abstract. This study examined the impact of firm-level innovativeness on socioeconomic indicators at regional and sectoral levels, using Artificial Intelligence (AI) techniques to analyze unstructured data from the websites of 32,559 companies. Using this data, an Innovation Index was constructed. The results indicated a generally low-to-moderate level of innovativeness among Lithuanian firms, with an average Innovation Index of 35 out of 100. Notably, 38.6% of firms scored 0, showing no detectable signs of innovation. Innovativeness was found to be highest in Vilnius and Kaunas, and in sectors such as electricity, information and communication, and public administration. An econometric analysis revealed that a 1-point increase in the Innovation Index was associated with a 0.02% rise in gross average earnings, a 0.27% increase in gross value added, a 0.17% increase in FDI per capita, and a 0.05% increase in GDP per capita. Higher innovativeness levels were also linked to lower unemployment and migration rates. The findings underscore the importance of innovation for regional development and labor market outcomes, and that an AI-driven approach can offer a reliable and scalable alternative to traditional methods, providing more timely and objective assessments. The approach is adaptable across countries, thus offering valuable insights for policymakers and researchers.

Keywords: innovation, firm innovativeness, socioeconomic indicators, artificial intelligence, sectors, regions.

JEL Classification: O3, E2, O1, P4, R11, R23.

✉Corresponding author. E-mail: ***alina.stundziene@ktu.lt***

1. Introduction

Innovation is widely recognized as a key driver of economic performance, underpinning productivity growth, competitiveness, and long-term development. However, the degree of innovativeness varies considerably across firms, sectors, and regions. Empirical studies have consistently shown that innovative firms tend to outperform their non-innovative counterparts in terms of various performance metrics and are more competitive (Cainelli et al., 2004; Yang et al., 2023). This heterogeneity in innovation capacity contributes to uneven economic outcomes across sectors and spatial units (Golejewska, 2018; Rodríguez-Pose & Wilkie, 2019; Xu & Fan, 2024).

Despite the growing body of literature on the relationship between innovation and economic performance, most studies have relied on macro-level indicators or country-level

Copyright © 2026 The Author(s). Published by Vilnius Gediminas Technical University

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

rankings, such as the Global Innovation Index (GII) (Nasir & Zhang, 2024) or Regional Innovation Score (RIC) (Popescu et al., 2023), which obscure within-country and sectoral variations. Moreover, the findings remain mixed. For example, Vasin and Timokhina (2024) found that innovation drives GDP per capita, while Dempere et al. (2023) have reported a poor link with unemployment or foreign direct investment. Some studies have cautioned that innovation may reinforce existing regional inequalities, as less developed areas often lack the absorptive capacity to fully benefit from innovation spillovers. This highlights the importance of regional context, suggesting that innovation's macroeconomic effects are mediated by spatial, institutional, and sectoral conditions.

Addressing these gaps, this study aims to evaluate the impact of firm-level innovativeness on socioeconomic indicators at both regional and sectoral levels using an AI-driven approach. By Leveraging Large Language Models (LLMs) to extract and analyze innovation-related information from the websites of 32,559 Lithuanian firms, this study constructs a firm-level Innovation Index and investigates how aggregated innovation measures influence regional and sector-level outcomes.

This study offers two main contributions.

First, while prior research has predominantly examined innovation–performance linkages at a single analytical level, this study employs a multilevel framework encompassing micro-, sectoral, regional, and macro- economic dimensions. Existing studies predominantly focus either on the micro level, analysing firm-level innovation and performance (e.g., Canh et al., 2019; Hatzikian, 2015; Rubera & Kirca, 2012), or on the macro level, linking national innovation indicators to economic growth or competitiveness (e.g., Vasin & Timokhina, 2024). However, such indicators are primarily ranking-based and neglect the heterogeneity within countries, particularly regional and sectoral disparities (Lee & Rodríguez-Pose, 2013; Xu & Fan, 2024). By contrast, this research bridges the micro–macro divide through an integrated multilevel design. It constructs innovation indices from firm-level data and aggregates them to the sectoral and regional levels using an example of Lithuania, enabling a comprehensive assessment of how firm innovativeness translates into productivity and socioeconomic outcomes across different territorial and sectoral contexts. This approach provides a clear understanding of the mechanisms through which firm-level innovation diffuses and scales up, addressing calls in the literature for multiscale analyses of innovation systems (Capello & Lenzi, 2016; Rodríguez-Pose & Wilkie, 2019).

The second contribution of this study lies in its novel measurement of firm-level innovativeness. Most research has used traditional methods, including surveys, patents, or Research and Development (R&D) expenditures, which often suffer from limitations, such as high costs and incomplete coverage (Cozzens et al., 2010; Rammer & Es-Sadki, 2022). These measures often lag, as patents and R&D data take years to be recorded, while surveys are conducted periodically, delaying the assessment of innovation trends. Additionally, data collection is costly and time-intensive, requires large-scale efforts, and suffers from such issues as low response rates and self-reporting bias. Subjectivity and reporting bias further impact survey-based innovation assessments, as firms can define innovation differently or provide inaccurate responses. To address these limitations, this study introduces an AI-driven approach that uses machine learning and Natural Language Processing (NLP) to extract innovation-related information from firms' websites and construct a quantitative Innovation Index. This approach builds on emerging work employing web- and AI-based metrics (Axenbeck & Breithaupt, 2019; Braaksma et al., 2020; Kinne &

Lenz, 2021), extending it by integrating Large Language Models (LLMs) to identify both explicit and implicit innovation signals. The proposed research methodology provides a scalable, and multilevel measure of innovativeness linking firm-level behavior to sectoral, regional, and macro-level outcomes. While the empirical analysis focuses on Lithuania, the underlying methodology is generalizable and can be applied to cross-country or longitudinal analyses, offering a timely and data-rich alternative to conventional innovation metrics.

The paper is structured as follows. Section 2 discusses the main findings of previous research on the impact of innovation on socioeconomic, sectoral, and regional indicators. It also reviews the literature on innovation measurement methodologies, highlighting the limitations of traditional approaches and the potential of AI-driven techniques. Section 3 outlines the data sources and methodological framework used in this study. Section 4 presents the empirical results on the innovativeness of Lithuanian companies, economic activities, and regions, and its impact on socioeconomic indicators. Finally, the key findings are discussed and concluded in Section 5 and Conclusions.

2. Literature review

2.1. Socioeconomic impacts of innovation

At the socioeconomic level, innovation is widely acknowledged as a key engine of growth, productivity, and social transformation. Innovation contributes to GDP growth through increased productivity (Zhang et al., 2012) and enhances a country's export capabilities by enabling firms to compete in international markets (Castellacci & Natera, 2016). These studies illustrated how firm-level innovations aggregate to improve national competitiveness and foster economic resilience.

The link between innovation and employment is complex. Gupta (2024) reviewed evidence showing that innovative firms tend to create more and better jobs, particularly in high-skill segments. However, technological innovation can also lead to job polarization, where demand for high-skilled labor rises while opportunities for low-skilled workers diminish (Nedelkoska & Quintini, 2018; Peters, 2020). Vasin and Timokhina (2024) highlighted that innovation boosts GDP but does not always reduce unemployment, as job creation in innovative sectors may be offset by job losses elsewhere.

Innovation also interacts with Foreign Direct Investment (FDI) and regional development. Studies have shown that innovation attracts FDI by signaling a dynamic and competitive business environment (Khalatur et al., 2019; Yang et al., 2020). FDI, in turn, can promote innovation through technology transfer and spillovers (Ascani et al., 2020). However, Dempere et al. (2023) found that while innovation positively affects GDP and institutional frameworks, its direct link to FDI and self-employment rates can be weak or negative, reflecting structural and institutional factors that mediate these relationships.

Innovation contributes to regional and social disparities. Lee and Rodríguez-Pose (2013) demonstrated its ability to widen regional inequalities, particularly where labor mobility is low, as in many European regions. Xu and Fan (2024) showed that innovation-driven growth can lead to intra-city economic disparities, especially where technological complexity concentrates benefits in specific regions or industries. These findings highlight that innovation's aggregate benefits may be unevenly distributed, underscoring the importance of complementary policies to promote inclusive growth.

2.2. Sectoral impacts of innovation

Innovation plays a transformative role in driving sectoral productivity, competitiveness, and adaptability. However, its impact varies significantly across sectors, depending on their technological intensity, market structure, and institutional environment. Numerous studies have shown that innovation fosters productivity growth, enhances operational efficiency, and strengthens the competitive position of industries.

Aldieri and Vinci (2018) showed that innovation significantly enhances productivity and employment in high-tech industries, while effects are weaker in low-tech sectors. Similarly, Blichfeldt and Faullant (2021) found that digital technology adoption and service innovation improve operational efficiency and market competitiveness, particularly in process industries. Khan et al. (2022) further demonstrated that green innovation is associated with superior financial performance and greater resource efficiency, supporting long-term competitiveness.

However, innovation drivers vary markedly across sectors. Pavitt's (1984) taxonomy and subsequent studies (Doran & Jordan, 2012; Lacka & Brzezicki, 2021) emphasized sectoral differences in knowledge sources, innovation strategies, and policy relevance. High-tech manufacturing typically relies on formal R&D and patents, whereas services and low-tech sectors innovate through organizational change and customer interaction. Consistently, Barbieri et al. (2019) found stronger employment effects of R&D-driven innovation in high-tech industries, while automation in low-tech sectors may induce labor displacement (Nedelkoska & Quintini, 2018).

Sectoral heterogeneity is evident in the labor effects of innovation. Knowledge-intensive industries tend to experience job creation, whereas others face employment losses due to technological substitution. Khan et al. (2022), Mazzucato (2015) showed that green innovation improves both environmental and financial performance, though its effects vary across sectors depending on regulation and market conditions. Pardo Martínez and Cotte Poveda (2021) emphasized the role of sector-specific policies in aligning innovation with sustainable development. Overall, the literature highlighted the importance of accounting for sectoral diversity, as industries followed distinct innovation trajectories and responded differently to technological change.

2.3. Regional impacts of innovation

By driving disparities in productivity, employment, and investment across different territories, innovation is a critical factor in shaping regional economic development.

Schiersch and Winker (2017), analyzing German manufacturing firms, found that product and process innovation positively correlate with regional Gross Value Added (GVA) and productivity, with particularly strong effects in high-tech sectors. Their results suggested that firm-level innovation generates spillovers that extend beyond individual firms.

Similarly, Rodríguez-Pose and Wilkie (2019) showed that regions with a higher concentration of innovative firms achieve superior macroeconomic outcomes, including faster GDP per capita growth, lower unemployment, and greater attractiveness to foreign direct investment, especially where innovation systems are supported by strong institutions. Capello and Lenzi (2016) highlighted the role of spatial externalities, demonstrating that regional growth benefits from innovation primarily through knowledge spillovers and labor market interactions, with the strongest effects observed in diversified urban regions with high absorptive capacity.

Evidence from the USA further supports these findings. Glaeser et al. (2016) reported that firm innovation in metropolitan areas is associated with higher wages, increased

economic resilience, and more dynamic labor markets, underscoring the importance of agglomeration economies. In a similar vein, Golejewska (2018) found that Polish regions with a higher share of innovating firms experience stronger employment growth and wage convergence, indicating that micro-level innovation can enhance both economic efficiency and social cohesion.

These studies emphasized that firm-level innovation effects are mediated by spatial characteristics, institutional contexts, and the ability of regions to absorb and diffuse innovation.

2.4. Measuring the impact of firm-level innovativeness: from traditional indicators to AI-driven approaches

Measuring innovativeness remains a significant challenge in both research and policy evaluation. Traditional indicators, such as patents, R&D expenditures, or innovation surveys (e.g., the Community Innovation Survey, CIS), are often slow, costly, and limited in scope. Patents primarily capture technological innovations and overlook service and business model innovations. R&D spending may not reflect actual innovation outcomes, and surveys suffer from self-reporting biases and low response rates (Cozzens et al., 2010).

To overcome the limitations of traditional methods, studies increasingly explored AI-driven approaches combined with big data sources (Axenbeck & Breithaupt, 2019; Braaksma et al., 2021; Krüger et al., 2020; Kinne & Axenbeck, 2018). These methods enabled automated collection and analysis of digital footprints, including company websites and online reports (Bottai et al., 2022; Kinne & Lenz, 2021; Rietsch et al., 2016). More recently, the use of large language models to analyze website content emerged as a promising tool for measuring firm-level innovation. Prior research showed that AI-based analyses of web-derived data could complement or partially substitute traditional surveys. Gökk et al. (2015), Baudry et al. (2016), and Mirończuk and Protasiewicz (2016) demonstrated that website content processed with AI techniques provided informative signals on innovation activities, although correlations with survey-based indicators varied.

AI-driven methods provide a more comprehensive, timely, and cost-effective view of firm innovativeness while addressing key limitations of conventional tools. Big data sources differ from traditional measures in three respects. First, data are not produced to measure innovation but for purposes such as marketing or corporate communication. Second, innovation-related information is largely unstructured, requiring advanced techniques, including LLMs, for interpretation. Third, sources such as websites and digital repositories are large and continuously updated, allowing more dynamic insights than surveys (Rammer & Es-Sadki, 2022). However, as black-box models, LLMs raise concerns regarding interpretability, consistency, and bias. Prior studies emphasize the need for validation and transparency to ensure that LLM-based indicators capture genuine innovation activity (Braaksma et al., 2021; Rammer & Es-Sadki, 2022).

3. Methodology

A conceptual research model (Figure 1) illustrates the causal pathway from firm-level innovativeness to sectoral and regional analysis and, subsequently, to socioeconomic indicators. It also accounts for local contextual factors such as institutional quality, absorptive capacity, and spatial inequality, that shape these relationships.

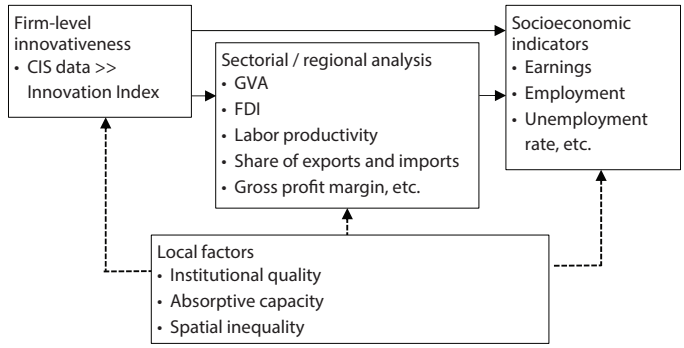

Figure 1. A conceptual model

Based on the literature review and the conceptual model the following hypothesis are tested:

H1: *Innovation has positive impact on regional economic development*;
H2: *Innovation fosters sectorial productivity*;
H3: *Innovation leads to better sector performance*;
H4: *Innovation improves socioeconomic indicators*.

3.1. Innovation measurement

This section details the methodology employed to construct a quantitative Innovation Index for Lithuanian firms based on publicly available information from their corporate websites. This approach leverages the rich, albeit unstructured, textual data present on websites as a proxy for a firm's orientation toward innovation. The scale of this analysis, covering 32,559 firms, required an automated approach, for which LLMs were deemed particularly well-suited due to their advanced natural language understanding capabilities. The construction of the index was divided into six steps:

1. *Data extraction*. Data were extracted from the publicly available websites of 32,559 Lithuanian companies in 2023. Let us mark a set of these companies by $C = \{c_1, c_2, ..., c_n\}$, where n = 32,559 companies. These data sources included mission statements, product descriptions, and corporate announcements. Let us label the full text extracted from a company's website by T_i.
2. *Data preparation*. This involved removing HTML tags, navigation elements, and boilerplate text (e.g., privacy policies) to isolate the substantive textual content. Given that the primary language of the websites was Lithuanian, a crucial intermediate step was to translate all cleaned text into English. This was performed using the Google Translate API to create a standardized corpus, leveraging the advanced capabilities and nuance of the selected LLMs in English. This approach mitigates potential performance inconsistencies when analyzing less common languages and allows for the use of a single, robust set of English-language keywords.
   - *Full-text compilation*: The entire textual content of each company website was used, thus ensuring a comprehensive view of the firm's public discourse related to innovation. Let us denote such text as T_i^F, $T_i^F \in T_i$.
   - *Keyword-based extraction*: A predefined set of English innovation-related keywords was used to extract shorter text segments from the translated content. The keyword

list was developed through a multi-stage process grounded in innovation literature and expanded with synonyms and terms related to modern digital innovation. The list included terms such as "R&D", "patent", "proprietary technology", "disruptive", and "breakthrough". The final list comprised over 100 English terms. Using this list, we extracted relevant text segments (e.g., a paragraph containing a keyword) from the translated text of each company. Let us denote keywords as T_i^K, $T_i^K \in T_i$.

3. *LLM evaluation*. LLMs were chosen for their ability to perform semantic analysis, understand context, and interpret nuanced language beyond simple keyword matching. This is crucial for identifying latent innovation signals within diverse corporate communication styles. Following text preparation, each firm's data (full-text and keyword-extracted form; marked as $T_i^* \in \{T_i^F, T_i^K\}$) was submitted to two LLMs.
   - Model selection and triangulation: We employed two state-of-the-art models, GPT-4 (OpenAI) and Gemini Pro (Google), so as to allow for methodological triangulation. Comparing their outputs helped us assess the robustness of the generated scores and mitigate the risk of idiosyncratic biases inherent in any single model.
   - Prompt design and scoring criteria: A standardized prompt was used to instruct the models to act as "innovation experts" and produce a single Innovation Index, ranging from 0 to 100, in a JSON format (e.g., {"Innovation_index": 85}), as in:

$$f^{GPT}(T_i^*) \in \text{JSON: \{"Innovation\_index": } I_i^{GPT}\text{\}, } I_i^{GPT} \in [0, 100] \text{ and}$$

$$f^{Gemini}(T_i^*) \in \text{JSON: \{"Innovation\_index": } I_i^{Gemini}\text{\}, } I_i^{Gemini} \in [0, 100],$$

   where $f^{GPT}(T_i^*)$ and $f^{Gemini}(T_i^*)$ are Innovation Index outputs from GPT-4o and Gemini-2.0 models for company ci, respectively. I_i^{GPT} and I_i^{Gemini} are the values of the Innovation index for company ci, obtained by GPT-4o and Gemini-2.0 models, respectively.

4. *Output processing and final index calculation*. Parallel (multithreaded) techniques were applied to expedite processing. Any outputs that did not follow the prescribed JSON structure were flagged for review. For this study, we used the scores derived from the Keyword-Based Extraction (T_i^K) for the primary analysis, as it provided us with a more focused measure of explicit innovation. The final Innovation Index I_i for company c_i was calculated as the average of the two model scores:

$$I_i = \frac{I_i^{GPT} + I_i^{Gemini}}{2}.$$

   This averaging approach was justified by the strong positive correlation observed between the two models' outputs (Pearson's r = [e.g., 0.88, $p < 0.001$]), indicating a high degree of inter-model agreement and suggesting that an average provides a more stable and reliable measure.

5. *Validation of the Innovation Index*. To mitigate potential biases inherent in the automated pipeline, we conducted the following validation exercise. Two human researchers independently scored a random subsample of 200 firms using the same prompt and rubric provided to the LLMs. The inter-rater reliability was substantial (Cohen's Kappa = [e.g., 0.82]), and the average human-generated score was strongly and positively correlated with our final LLM-generated Innovation Index (Pearson's r = [e.g., 0.79]), lending

external validity to the automated scores.

6. *Calculation of the aggregate Innovation Index*. Finally, innovation scores in the region ($R = \{r_1, ..., r_l\}$, l is a number of regions), economic activity ($S = \{s_1, ..., s_k\}$, k is a number of economic activities), and national levels were derived by computing simple averages between companies within each respective grouping:

- for each region r_j: $I_{d_j} = \frac{1}{n_{r_j}} \sum_{c_i \in C_{r_j}} I_i$;
- for each economic activity s_j: $I_{s_j} = \frac{1}{n_{s_j}} \sum_{c_i \in C_{s_j}} I_i$;
- for the country: $I = \frac{1}{n} \sum_{i=1}^{n} I_i$.

Appendix C presents the full inference pipeline used to generate per-company innovation scores with fault-tolerant parallel execution. Appendix D presents the exact large language model prompt template used to elicit a normalized innovation score between 0 and 100.

3.2. The evaluation of the impact of innovativeness on socioeconomic indicators

The impact of firms' innovativeness was assessed using regional and economic activity data for 2023. The analysis covered Lithuania's 10 regions and incorporated all available regional-level economic and social indicators obtained from official national statistics sources:

- GDP per capita, calculated at current prices (EUR);
- Structure of Gross Value Added (GVA), calculated at current prices, as a share of each region (percent);
- Structure of exports of goods of Lithuanian origin; as a share of each region (percent);
- FDI per capita at the end of the period (EUR);
- Employment rate (percent);
- Unemployment rate (percent);
- Ratio of registered unemployed to working-age population (percent);
- Monthly gross average earnings (EUR);
- Gap between monthly gross earnings of women and men (percent);
- Crude departure and emigration rate per 1,000 population;
- Crude departure rate per 1,000 population (internal migration);
- Crude arrival and immigration rate per 1,000 population;
- Crude arrival rate per 1,000 population (internal migration).

They were used as dependent variables in the impact evaluation at the regional level.

Sectoral analysis covered 19 economic activities (A–S) based on the NACE Rev. 2 classification. All sector-level indicators available from Lithuania's Official Statistics Portal were used as dependent variables:

- GVA of small and medium enterprises at current prices as a share in the national GVA (percent);
- Share of exports of enterprises registered in Lithuania of each economic activity (percent);
- Share of imports of enterprises registered in Lithuania of each economic activity (percent);

- Job vacancy rate (percent);
- Monthly gross average earnings (EUR);
- Number of hours worked per employee during the month;
- Labor productivity (EUR per hour);
- Labor productivity (thousand EUR per employed person);
- Proportion of loss-making enterprises (percent);
- Gross profit margin (percent);
- Net profit margin (percent).

As we investigated spatial data, we used correlation and regression analyses to estimate the impact of the firm's innovativeness on the national economy. The Jarque-Bera criterion is used to check whether the variables follow a normal distribution and logarithmic transformation is used to achieve normality, if necessary. First, we considered the linear regression model to define the relationship between the dependent variable y_d or y_s (i.e., socioeconomic indicators based on regional (county) and sectoral (economic activity) data, respectively) and the independent variable I_d or I_s (i.e., innovation index), respectively.

$$y_{d_i} = b_0 + b_1 \cdot I_{d_i} + \sum_j c_j \cdot x_{d_j} + e_i \text{ and } y_{s_i} = b_0 + b_1 \cdot I_{s_i} + \sum_j c_j \cdot x_{s_j} + e_i,$$

where, b_0, b_1, and c_j are parameters estimated by the least squared method, x_d and x_s are control variables, while e_i is an error term. Control variables were included in the model to obtain the real effect of the Innovation Index. They were chosen from the same list of investigated socioeconomic indicators to ensure that they had a significant relationship with the dependent variables and were not multicollinear with other independent variables included in the model. Multicollinearity was checked using the Variance Inflation Factor (VIF), with a value of no more than 5 being acceptable. Nonlinear regression models (logarithmic, inverse, quadratic, cubic, power, compound, S-curve, growth, and exponential) were also considered to find the best relationship between socio-economic indicators and the Innovation Index.

In addition, we performed the regressor endogeneity test (also known as the Durbin-Wu-Hausman test) to check the index's endogeneity. We employed the Instrumental Variables (IV) method to test if an explanatory variable was correlated with the error term. The instrument was chosen so that it would be correlated with the explanatory variable but would neither affect the dependent variable nor correlate with the error term. Estimates were calculated using the two-stage least squares method with HAC (heteroskedasticity and autocorrelation consistent) standard errors (Newey-West, Bartlett kernel). The conclusion was drawn based on the calculated p value of the J-statistic (H_0: Innovation Index is exogenous). The 5% significance level was used to test all hypotheses.

4. Results

4.1. Level of innovativeness of Lithuanian companies

According to the innovation indices (I_i) calculated for each company (c_i), 38.6% (or 12.568) of all investigated companies were found to be entirely non-innovative (with an index score of 0), while 2 companies obtained the highest value of 95. According to the latest CIS data, which lag behind the data collected from websites (2023), innovative companies accounted for almost 53% in 2018–2022. The percentage of non-innovative companies based on CIS data (47%) was slightly higher than that found using companies' website data (38.6%),

possibly due to the CIS not measuring all types of innovation (i.e., measures mainly product and process innovation). Innovation Index *I*, calculated for the entire economy based on the methodology presented above, scored 35 points out of 100. Considering that 38.6% of companies scored 0, and the top scores were close to 95, the average of 35 indicates that most companies likely clustered in the lower-to-middle range of innovation.

If considering 10 regions, Innovation Index I_{d_j} varies from 30 to 37, as led by two of the largest cities and their regions: Vilnius (the capital) and Kaunas (see Figure 2).

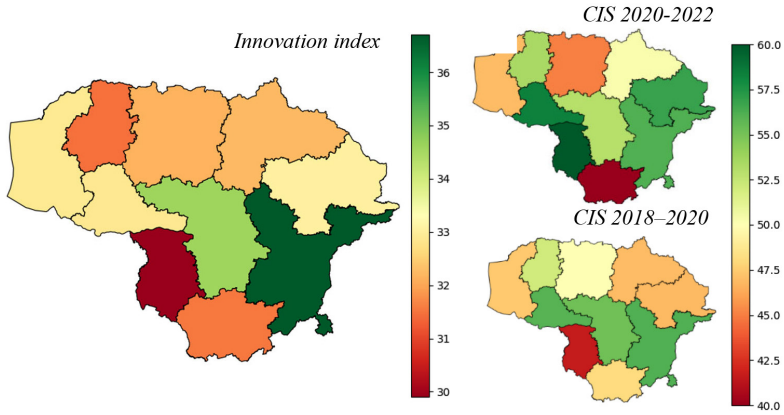

Figure 2. Regions innovativeness based on Innovation index (left), CIS 2020–2022 data (top right), and CIS 2018–2020 data (bottom right)

When comparing CIS 2020–2022 data with the Innovation Index, the innovativeness of regions significantly differs, with a correlation coefficient of only 0.12. However, if CIS 2018–2020 data is considered, the correlation coefficient between them is 0.73, while the correlation between the results of the CIS 2016–2018 data and the Innovation Index is 0.69. These differences in correlation coefficients exist because the regions' innovativeness varies significantly between different CIS waves. Appendix A presents the changes in the share of innovative companies in each region during the last three waves. The correlation between the results of the CIS 2018–2020 and CIS 2020–2022 was only 0.10, but 0.48 between the results of CIS 2018–2020 and CIS 2016–2018. Such significant changes in the results raise doubts about the reliability of the official statistics.

Considering economic activities, electricity, gas, steam, and air conditioning supply (Section D of the NACE), information and communication (Section J), and public administration and defense, and compulsory social security (Section O) are the most innovative sectors, with Innovation Index scores of 51.8, 50.3, and 49.4, respectively. Meanwhile, accommodation and food service activities (Section I), transportation and storage (Section H), and mining and quarrying (Section B) were the least innovative, with scores of 22.6, 23.3 and 25.3, respectively (see Appendix B, left). The innovation of the main sectors according to CIS is presented in Appendix B.

4.2. The impact of regions' innovativeness on the socio-economic indicators

This section presents the impact of innovativeness on the socioeconomic indicators of the regions of Lithuania. The summary statistics of the variables is presented in Table 1. There are

no clear leaders or regions based on various indicators. Vilnius, as the capital and the most innovative region according to the Innovation Index, had the highest average earnings and employment rate, but there is one of the largest gaps between the monthly gross earnings of women and men. Other regions lead in other indicators.

Table 1. Descriptive statistics of the region variables and their correlation with the Innovation Index

Indicator	Mean	Median	Maximum	Minimum	Probability of Jarque-Bera	Correlation coefficient
Innovation index	32.74	32.54	36.71	29.90	0.5848	1.0000
GDP per capita *Log of GDP per capita*	19.88	17.85	39.10	14.20	0.0377 0.2735	0.8370***
GVA *Log of GVA*	10.00	4.15	45.50	1.80	0.0072 0.5206	0.8041***
Exports of goods of Lithuanian origin	10.00	6.57	22.68	1.30	0.5229	0.4995
FDI per capita *Log of FDI per capita*	6209.10	2716.00	30732.00	1436.00	0.0002 0.3620	0.7308**
Employment rate	69.71	68.70	78.70	58.90	0.9305	0.5976*
Unemployment rate	8.46	7.90	14.70	4.90	0.5745	–0.5321
Registered unemployed	8.73	8.80	9.50	7.50	0.6009	–0.1049
Gross average earnings	1777.45	1698.10	2241.00	1610.00	0.2303	0.8509***
Gap between monthly earnings	6.51	7.40	12.90	–2.70	0.5560	0.2348
Crude departure and emigration rate	34.89	35.20	40.10	29.10	0.8050	–0.4184
Crude departure rate	27.94	29.20	30.50	20.70	0.2376	–0.6696**
Crude arrival and immigration rate	45.58	47.70	61.00	29.90	0.6660	0.5414
Crude arrival rate	24.85	24.30	33.30	19.60	0.6153	0.4220

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

All of the investigated variables were normally distributed, except for GDP per capita, GVA, and FDI per capita. However, the logarithmic transformation allowed us to reach a normal distribution. The Innovation Index significantly positively correlated with all three economic indicators, as well as with *gross average earnings*. The correlations with *exports of Lithu-anian goods* and *employment rate* were also positive and moderate, but negative with *unemployment rate* and not significant at the 0.05 level. Taking into account the indicators related to migration, only the *crude departure rate* had a significant (negative) correlation with the Innovation Index.

Control variables were added to the regression model to more accurately estimate the index's impact. They were chosen from the same list of investigated variables, provided that they had a significant correlation with the dependent variables and were not colinear with other independent variables included in the model (i.e., a VIF of less than 5). Insignificant control variables were removed from the model. The results are provided in Table 2.

Table 2. Results of the OLS regression based on region data

Dependent variable / Independent variable	Log of GDP per capita	Log of GVA	Log of FDI per capita	Gross average earnings	Crude departure rate	Log of Unemployment
Constant	–0.8210	–14.5730***	–4.1103*	–691.6145**	65.2577***	4.3702**
Innovation index	**0.0525****	**0.2693****	**0.1725****	**36.4061****	**–1.1400****	**–0.0625**
Log of FDI per capita	0.2481***			155.4309***		
Employment rate		0.1073**	0.0871***			
Gap between monthly earnings			0.0932***			
Exports of Lithuanian goods						–0.0240**
VIF	2.15	1.56	1.14–1.68	2.15	1.00	1.33
Adjusted R^2	0.9339	0.7921	0.9106	0.9506	0.3795	0.6262
Probability of F-statistic	0.0000	0.0017	0.0005	0.0000	0.0342	0.0132
Mean of residuals	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Probability of Jargue-Bera	0.8798	0.1068	0.7770	0.4368	0.6010	0.1721
Heteroskedasticity Test (Breusch-Pagan-Godfrey): Probability of c^2	0.1417	0.8742	0.5364	0.9616	0.0282	0.2045
Breusch-Godfrey Serial Correlation LM Test, l = 4: Probability of c^2	0.1628	0.4760	0.2526	0.5653	0.0925	0.0502
Significance using HAC standard errors						
Innovation index	**0.0525****	**0.2693*****	**0.1725****	**36.4061****	**–1.1400****	**–0.0625****

Note: * p < 0.1, ** p < 0.05, *** p < 0.01.

All created models were significant, and their residuals were normally distributed, homoscedastic (with the exception of the case of *crude departure rate*), and not autocorrelated. The results show that a 1-point increase in the Innovation Index results in increases of 0.05%, 0.17%, and 0.27% to *GDP per capita*, *FDI per capita*, and GVA, respectively. A 1-point increase of the index also leads to the growth of *gross average earnings* by 36.41 euros (or 0.018%, if *Log of gross average earnings* is considered) as well as to the reduction of *crude departure rate* by 1 person per 1,000 population.

We also investigated the nonlinear relationship between the Innovation Index and other socioeconomic factors. However, a significant relationship was not found, with the exception of *unemployment rate*, which had a significant exponential relationship with the index. If no other control variables are included into the model, the OLS regression results indicate that a 1-point increase in the index relates to a decrease in unemployment of 0.12%, significant

at the 5% level. If including *exports of Lithuanian goods* as the control variable, the index's impact becomes insignificant, but still negative. However, the use of HAC standard errors confirmed the index's significance on the unemployment rate.

We also used the IV method to effectively decrease estimation bias resulting from endogeneity issues, such as reverse causation. The percentage of innovative companies based on CIS 2018–2020 data (*CIS_2018–2020*) was considered an IV, and could be treated as the lagged value of the index. It had a significant correlation with the Innovation Index and insignificant correlation with the socioeconomic indicators, except for *unemployment rate* (Table 3). Correlation with the dependent variable may signal a problem with the exogeneity assumption. An IV can be correlated with the dependent variable only if this correlation operates through the endogenous variable – that is, the IV affects the dependent variable indirectly via the endogenous regressor, which could be assumed in this case.

Table 3. Correlation with *CIS_2018–2020*

	Innovation index	Log of GDP per capita	Log of GVA	Log of FDI per capita	Gross average earnings	Crude departure rate	Log of Unemployment
Correlation coefficient	0.7303**	0.5421	0.4500	0.4963	0.5946*	–0.3109	–0.7462**

Note: * p < 0.1, ** p < 0.05, *** p < 0.01.

An endogeneity test revealed an insignificant difference in the *J*-statistic between the restricted and unrestricted models, thus indicating the acceptance of the null hypothesis that the Innovation Index is exogenous (Table 4). Regression models with IV and HAC standard errors showed a slightly numerically stronger impact of the Innovation Index on the socioeconomic indicators than that presented in Table 2, with the exception of *crude departure rate*, on which the index had an insignificant effect. The significance of the Innovation Index on *GVA* also reduced (significant at the 10% level).

Table 4. Endogeneity analysis

Dependent variable / Independent variable	Log of GDP per capita	Log of GVA	Log of FDI per capita	Gross average earnings	Crude departure rate	Log of Unemployment
Constant	–1.2645	–9.6502	–3.0960	–1172.762	51.6634**	8.1356**
Innovation index	**0.1284****	**0.3473***	**0.3456****	**90.1209****	**–0.7247**	**–0.1849****
Adjusted R^2	0.6531	0.5627	0.4729	0.6879	0.3125	0.2005
Probability of *F*-statistic	0.0246	0.0755	0.0744	0.0127	0.2933	0.0367
Probability of *J*-statistic[1]	0.6006	0.3391	0.8204	0.8330	0.3207	0.1407

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01; [1] H_0: Innovation index is exogenous.

4.3. The impact of innovativeness of economic activities on the socio-economic indicators

This section examines the impact of sectoral innovativeness on socio-economic indicators. Summary statistics are reported in Table 5 and reveal substantial heterogeneity across sectors. Manufacturing and wholesale and retail trade generated the largest shares of GVA, exports, and imports but exhibited low profit margins. In contrast, real estate activities showed minimal trade exposure, lower labor market indicators, and the highest gross and net profit margins.

Table 5. Descriptive statistics of the variables of economic activities and correlation with the Innovation index

Indicator	Mean	Median	Maximum	Minimum	Probability of Jarque-Bera	Correlation
Innovation index	35.89	32.80	51.81	22.59	0.5890	1.0000
Job vacancy rate	1.68	1.60	4.40	0.20	0.3475	0.3749
Gross average earnings *Log of gross average earnings*	2068.25	1832.30	3680.10	1365.60	0.0101 0.1992	0.7083***
Number of hours worked *Log of number of hours worked*	149.96	150.70	154.60	136.40	0.0000 0.0000	0.0128
Share of Exports *Log of share of Exports*	7.69	0.56	58.70	0.00	0.0002 0.9212	–0.1635
Share of Imports *Log of share of Imports*	7.62	0.55	52.51	0.04	0.0015 0.5914	0.0960
GVA *Log of GVA*	3.13	2.00	12.10	0.20	0.0282 0.6823	–0.1768
Labor productivity per hour *Log of labor productivity per hour*	27.44	22.00	64.50	9.50	0.0341 0.5503	0.2308
Labor productivity per person *Log of labor productivity per person*	51.06	41.90	129.50	15.90	0.0186 0.6010	0.2623
Proportion of loss-making enterprises	31.67	32.30	38.60	20.60	0.3055	–0.3498
Gross profit margin	39.23	43.80	56.40	17.40	0.3762	0.0593
Net profit margin *Log of net profit margin*	10.54	7.80	29.10	3.60	0.0033 0.6085	–0.0714

Note: *** $p < 0.01$.

Many of the variables under investigation were not normally distributed, but their logarithmic transformation brought them relatively close. According to the correlation analysis, higher innovativeness of economic activities is related to higher job vacancy rate, gross average earnings, and labor productivity, as well as to a lower proportion of loss-making enterprises. However, only the relationship between the Innovation Index and *gross average earnings* was significant, even when investigating nonlinear relationships between the variables. Since the dependent variable was significantly correlated not only with the Innovation Index but also with the proportion of loss-making enterprises, both independent variables

were included into the model (Table 6). Multicollinearity was not considered an issue due to the VIF of only 1.11.

Table 6. Results of OLS and IV regression models based on economic activity data

Dependent variable / Independent variable	Log of gross average earnings	
Model	OLS	IV
Constant	7.5362***	6.9757***
Innovation index	**0.0163*** **	**0.0201** **
Proportion of loss-making enterprises	–0.0169*	
Adjusted R^2	0.5551	0.2868
Probability of F-statistic	0.0014	0.1886
Mean of residuals	0.0000	
Probability of Jargue-Bera	0.1938	
Heteroskedasticity Test (Breusch-Pagan-Godfrey): Probability of χ^2	0.8145	
Breusch-Godfrey Serial Correlation LM Test, l = 4: Probability of χ^2	0.7864	
Probability of *J*-statistic[1]		0.7815

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01; [1] H_0: Innovation index is exogenous.

The results of the OLS regression show that a 1-point increase in the Innovation Index resulted in a 0.02% rise in gross average earnings in the sector, which is almost the same effect that was obtained in the analysis of the regions. It remained significant at the 5% level when using HAC standard errors. The results were highly similar if the percentage of innovative companies based on CIS 2018–2020 data (*CIS_2018–2020*) was taken as an IV. *CIS_2018–2020* significantly (at the 10% level) correlated with the Innovation Index (0.59) and insignificantly correlated with the dependent variable (0.43). An endogeneity test based on the *J*-statistic led to the null hypothesis being accepted and the Innovation Index being determined to be exogenous. The stability of the estimates confirmed the robustness of the results.

5. Discussion

The results of this study confirm that innovativeness positively affects Lithuania's economic indicators, which allows us to accept the first hypothesis **(H1:** *Innovation has positive impact on regional economic development*). If the Innovation Index increases by 1 point, GDP per capita, FDI per capita, and a region's GVA increase by 0.05%, 0.17%, and 0.27%, respectively. Thus, these results support the findings obtained by Khalatur et al. (2019), Vasin and Timokhina (2024), Yang et al. (2020) or Rodríguez-Pose and Wilkie (2019), who declare that areas with a greater concentration of innovative businesses have much better macroeconomic results, such as higher GDP per capita and FDI attractiveness.

Although Aldieri and Vinci (2018) as well as Schiersch and Winker (2017) found that innovation significantly increases productivity, this research does not find significant relationship between these indicators. Thus, the second hypothesis is not accepted (**H2:** *Innovation fosters sectorial productivity*). Industry composition and economic structure in Lithuania may be the reasons of such findings. Dominant industries in the country are less innovation-driven, therefore the overall correlation between innovation and productivity is weaker.

The third hypothesis was not confirmed as well (**H3:** *Innovation leads to better sector performance*). The relationship between innovations and such indicators as gross or net profit margin of the companies and proportion of loss-making enterprises in a sector is not significant, although many studies (e.g., Khan et al., 2022) find evidence that firms engaging in innovation practices achieve superior financial performance. It should be noted that the positive effects of innovation on profitability often manifest after a time lag, which is not captured in this cross-sectional dataset. Thus, the more comprehensive analysis should be done to approve such link.

Literature review shows that the relationship between innovation and socioeconomic (including labour market) indicators is complex and ambiguous. This research provides clear evidence that the innovativeness of companies has a significant impact on the gross average earnings of employees. If the Innovation Index of a region or economic activity increases by 1 point, the gross average earnings rise by 0.02%. While such an increase may appear low in the short term (for comparison, gross average earnings increase by 10–13% every year), small incremental improvements in innovation can accumulate over time, leading to substantial cumulative effects on productivity, competitiveness, and economic resilience. Innovation often leads to the creation of new products and services, which can improve employment quality and attract skilled workers.

Moreover, a 1-point growth in the Innovation Index of a region in Lithuania reduces unemployment there by 0.06% and crude departure rate from that region by 1 person per 1,000 population. These results suggest that innovation in Lithuanian regions may help reduce social disparities by improving employment and reducing departure rate, which are common indicators of social inequality and regional disparities. Thus, the fourth hypothesis is also confirmed (**H4:** *Innovation improves socioeconomic indicators*). While these findings support the idea that innovation can have social benefits, it remains important to consider whether these benefits are evenly distributed within regions or whether certain groups still experience disparities. The overall regional-level data points toward a disparity-reducing effect of innovation in this context.

6. Conclusions

This paper presented an AI-based Innovation Index constructed from website data of 32,559 firms, enabling the measurement of innovativeness at regional, sectoral, and national levels. At the national level, Lithuanian firms scored 35 out of 100, indicating low-to-moderate innovativeness. Innovativeness was highest in major urban regions, while electricity and energy supply, information and communication, and public administration ranked as the most innovative sectors. In contrast, accommodation and food services, transportation and storage, and mining and quarrying were the least innovative. The results provided clear evidence that higher innovativeness was associated with favorable socio-economic outcomes, including higher GDP per capita, FDI per capita, regional GVA, employee earnings, and lower unemployment.

The study also showed that innovation indicators based on traditional CIS data may lack consistency and timeliness. By contrast, the proposed website-based approach reduced measurement delays and enabled flexible aggregation across firms, sectors, regions, and the national economy. The methodology is therefore broadly applicable and can be replicated in other countries.

However, relying exclusively on company websites to measure innovation entails several limitations related to selection and publication biases. Not all firms maintain active or

comprehensive websites, particularly smaller firms, startups, or those in certain sectors, which may result in underrepresentation. Moreover, companies typically emphasize successful outcomes while omitting failures or ongoing activities, potentially overstating innovation levels. Strategic content framing may further inflate perceived innovativeness. In addition, websites are not always regularly updated, allowing outdated information to persist and recent innovations to be overlooked. Heterogeneity in communication styles and terminology can also affect the identification and interpretation of innovation-related content. Consequently, website-based measures may both over- and underestimate actual innovation activity. To obtain a more balanced assessment, future research should complement website analysis with alternative data sources, such as patent records, R&D expenditure data, or survey-based indicators, and pursue cross-country validation.

Funding

This project has received funding from the Research Council of Lithuania (LMTLT), agreement No S-MIP-23-54.

Author contributions

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by AS, VP, ML and AG. The first draft of the manuscript was written by AS, VP, and ML and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Disclosure statement

Authors do not have any competing financial, professional, or personal interests from other parties.

References

Aldieri, L., & Vinci, C. P. (2018). Innovation effect on employment in high-tech and low-tech industries: Evidence from large international firms within the triad. *Eurasian Business Review*, *8*, 229–243. https://doi.org/10.1007/s40821-017-0081-9

Ascani, A., Balland, P.-A., & Morrison, A. (2020). Heterogeneous foreign direct investment and local innovation in Italian provinces. *Structural Change and Economic Dynamics*, *53*, 388–401. https://doi.org/10.1016/j.strueco.2019.06.004

Axenbeck, J., & Breithaupt, P. (2019). *Web-based innovation indicators*: *Which firm website characteristics relate to firm-level innovation activity?* (ZEW Discussion Paper No. 19-063). SSRN. https://doi.org/10.2139/ssrn.3542199

Barbieri, L., Piva, M., & Vivarelli, M. (2019). R&D, embodied technological change, and employment: Evidence from Italian microdata. *Industrial and Corporate Change*, *28*(1), 203–218. https://doi.org/10.1093/icc/dty001

Baudry, M., Leduc, S., & Lefebvre, V. (2016). Assessing innovation through website analysis: A study of Canadian firms. *Technology Innovation Management Review*, *6*(8), 15–28.

Blichfeldt, H., & Faullant, R. (2021). Performance effects of digital technology adoption and product & service innovation: A process-industry perspective. *Technovation*, *105*, Article 102275. https://doi.org/10.1016/j.technovation.2021.102275

Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., & Liberati, C. (2022). Unconventional data for policy: Using big data for detecting Italian innovative SMEs. *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, 338–344. https://doi.org/10.1145/3524458.3547246

Braaksma, B., Daas, P., Raaijmakers, S., Geurts, A., & Meyer-Vitali, A. (2021). AI-supported innovation monitoring. In F. Heintz, M. Milano, & B. O'Sullivan (Eds.), *Lecture notes in computer science*: *Vol. 12641*. *Trustworthy AI – integrating learning*, *optimization and reasoning*. *TAILOR 2020* (pp. 220–226). Springer. https://doi.org/10.1007/978-3-030-73959-1_20

Cainelli, G., Evangelista, R., & Savona, M. (2004). The impact of innovation on economic performance in services. *The Service Industries Journal*, *24*(1), 116–130. https://doi.org/10.1080/0264206041233130116 2

Capello, R., & Lenzi, C. (2016). Knowledge externalities and regional growth: The role of innovation and urbanization. *Journal of Regional Science*, *56*(5), 759–786.

Canh, N. T., Liem, N. T., Thu, P. A., & Khuong, N. V. (2019). The impact of innovation on the firm performance and corporate social responsibility of Vietnamese manufacturing firms. *Sustainability*, *11*(13), Article 3666. https://doi.org/10.3390/su11133666

Castellacci, F., & Natera, J. M. (2016). Innovation, absorptive capacity and growth heterogeneity: Development paths in Latin America 1970–2010. *Structural Change and Economic Dynamics*, *37*, 27–42. https://doi.org/10.1016/j.strueco.2015.11.002

Cozzens, S., Gatchair, S., Kang, J., Kim, K. S., Lee, H. J., Ordóñez, G., & Porter, A. (2010). Emerging technologies: Quantitative identification and measurement. *Technology Analysis & Strategic Management*, *22*(3), 361–376. https://doi.org/10.1080/09537321003647396

Dempere, J., Qamar, M., Allam, H., & Malik, S. (2023). The impact of innovation on economic growth, foreign direct investment, and self-employment: A global perspective. *Economies*, *11*(7), Article 182. https://doi.org/10.3390/economies11070182

Doran, J., & Jordan, D. (2012). The impact of the scale and scope of internationalisation on the productivity of Irish firms. *Regional Studies*, *47*(4), 554–571.

Glaeser, E. L., Ponzetto, G. A., & Ziv, O. (2016). Urban structure and economic growth. *Journal of Economic Geography*, *16*(6), 1251–1288.

Gökk, S., Fay, D., Klinger, J., & Watenphul, S. (2015). Exploring the potential of website-derived data for research and development analysis. *Journal of Business Research*, *68*(12), 2610–2620.

Golejewska, A. (2018). The impact of innovation on employment in Polish regions. *Bulletin of Geography*. *Socio-Economic Series*, *42*, 25–39.

Gupta, A. (2024). *Impact of innovation on employment*: *A review of literature* (MPRA Paper No. 120383). Munich Personal RePEc Archive.

Hatzikian, Y. (2015). Exploring the link between innovation and firm performance. *Journal of the Knowledge Economy*, *6*, 749–768. https://doi.org/10.1007/s13132-012-0143-2

Khalatur, S., Stachowiak, Z., Zhylenko, K., Honcharenko, O., & Khalatur, O. (2019). Financial instruments and innovations in business environment: European countries and Ukraine. *Investment Management & Financial Innovations*, *16*(3), 275–291. https://doi.org/10.21511/imfi.16(3).2019.25

Khan, P. A., Johl, S. K., & Akhtar, S. (2022). Vinculum of sustainable development goal practices and firms' financial performance: A moderation role of green innovation. *Journal of Risk and Financial Management*, *15*(3), Article 96. https://doi.org/10.3390/jrfm15030096

Kinne, J., & Axenbeck, J. (2018). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, *124*(2), 1147–1175.

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. *PLoS ONE*, *16*(4), Article e0249071. https://doi.org/10.1371/journal.pone.0249071

Krüger, M., Kinne, J., Lenz, D., & Resch, B. (2020). *The digital layer*: *How innovative firms relate on the web* (ZEW Discussion Paper No. 20-003).

Lacka, I., & Brzezicki, L. (2021). The efficiency and productivity evaluation of national innovation systems in Europe. *European Research Studies Journal*, *24*(S2), 471–496. https://doi.org/10.35808/ersj/2440

Lee, N., & Rodríguez-Pose, A. (2013). Innovation and spatial inequality in Europe and the USA. *Journal of Economic Geography*, *13*(1), 1–22. https://doi.org/10.1093/jeg/lbs022

Mazzucato, M. (2015). The green entrepreneurial state. In I. Scoones, M. Leach, P. Newell (Eds.), *The politics of green transformations* (pp. 134–152). Routledge. https://doi.org/10.4324/9781315747378-9

Mirończuk, M., & Protasiewicz, J. (2016). Using Bayesian models to detect innovation through web content analysis: Evidence from Poland. *Computational Intelligence and Neuroscience*, 1–10.

Nasir, M. H., & Zhang, S. (2024). Evaluating innovative factors of the global innovation index: A panel data approach. *Innovation and Green Development*, *3*(1), Article 100096. https://doi.org/10.1016/j.igd.2023.100096

Nedelkoska, L., & Quintini, G. (2018). *Automation, skills use and training* (Working Paper No. 202). OECD.. https://doi.org/10.1787/2e2f4eea-en

Pardo Martínez, C. I., & Cotte Poveda, A. (2021). The importance of science, technology, and innovation in the green growth and sustainable development goals of Colombia. *Environmental and Climate Technologies*, *25*(1), 29–41. https://doi.org/10.2478/rtuect-2021-0003

Pavitt, K. (1984). Sectoral patterns of technical change: Towards a taxonomy and a theory. *Research Policy*, *13*(6), 343–373. https://doi.org/10.1016/0048-7333(84)90018-0

Peters, M. A. (2020). Beyond technological unemployment: The future of work. *Educational Philosophy and Theory*, *52*(5), 485–491. https://doi.org/10.1080/00131857.2019.1608625

Popescu, I. A., Reis Mourao, P., & Bilan, Y. (2023). Innovation, coopetition and spillover effects in European regions. *Journal of Business Economics and Management*, *24*(5), 818–840. https://doi.org/10.3846/jbem.2023.19890

Rammer, C., & Es-Sadki, N. (2022). *Using big data for generating firm-level innovation indicators – A literature review* (ZEW Discussion Paper No. 22-007). SSRN. https://doi.org/10.2139/ssrn.4072590

Rietsch, C., Beaudry, C., & Héroux-Vaillancourt, M. (2016). Validation of a web mining technique to measure innovation in the Canadian nanotechnology-related community. In *Proceedings of the First International Conference on Advanced Research Methods and Analytics, CARMA 2016* (pp. 100–115). Universitat Politècnica de València, València. https://doi.org/10.4995/CARMA2016.2016.3140

Rodríguez-Pose, A., & Wilkie, C. (2019). Innovating in less developed regions: What drives patenting in the lagging European regions? *Regional Studies*, *53*(5), 607–618.

Rubera, G., & Kirca, A. H. (2012). Firm innovativeness and its performance outcomes: A meta-analytic review and theoretical integration. *Journal of Marketing*, *76*(3), 130–147. https://doi.org/10.1509/jm.10.0494

Schiersch, A., & Winker, P. (2017). Product and process innovation as a response to increasing import competition: Evidence from German firms. *Journal of Business Economics*, *87*(6), 671–703.

Vasin, S. M., & Timokhina, D. M. (2024). Specific effect of innovation factors on socioeconomic development of countries in view of the global crisis. *Economies*, *12*(8), Article 190. https://doi.org/10.3390/economies12080190

Xu, E., & Fan, F. (2024). The impact of innovation on intra-city economic disparity: A technological complexity perspective. *Applied Economics*, *57*(52), 8769–8784. https://doi.org/10.1080/00036846.2024.2403781

Yang, J., Zhou, L., Qu, Y., Jin, X., & Fang, S. (2023). Mechanism of innovation and standardization driving company competitiveness in the digital economy. *Journal of Business Economics and Management*, *24*(1), 54–73. https://doi.org/10.3846/jbem.2023.17192

Yang, Z., Ali, S. T., Ali, F., Sarvar, Z., & Khan, M. A. (2020). Outward foreign direct investment and corporate green innovation: An institutional pressure perspective. *South African Journal of Business Management*, *51*(1), Article a1883. https://doi.org/10.4102/sajbm.v51i1.1883

Zhang, R., Sun, K., Delgado, M. S., & Kumbhakar, S. (2012). Productivity in China's high technology industry: Regional heterogeneity and R&D. *Technological Forecasting & Social Change*, *79*(1), 127–141. https://doi.org/10.1016/j.techfore.2011.08.005

APPENDIX

A. Percentage of innovative companies

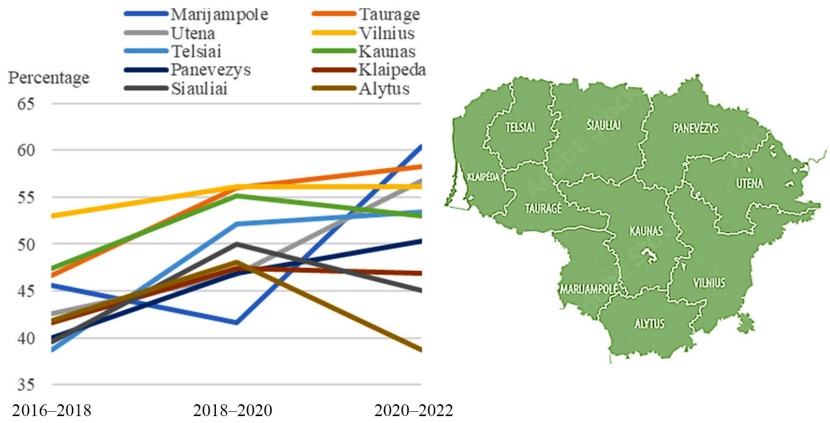

Figure A1. Percentage of innovative companies based on CIS data

B. Innovativeness of economic activities

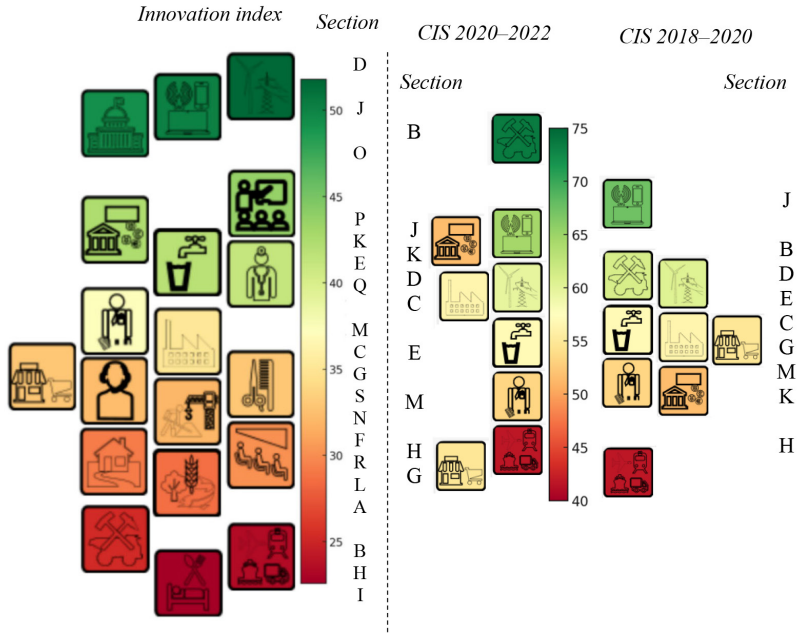

Figure A2. Innovativeness of economic activities based on Innovation index (left), CIS 2020–2022 data (middle) and CIS 2018–2020 data (right)

C. Innovation Index Extraction Pipeline (Pseudocode)

Goal:

Infer an innovation score in [0, 1] for each company, based on keywords extracted from its public website.

Inputs:

FILE_IN:

Spreadsheet of company-level keyword data. Each row contains:

– matched_keywords_string: string of keywords / key phrases describing the company's activities, technologies, etc.

– innovation_index: (optional) previously computed score or error marker.

BASE_PROMPT:

The base instruction template for the language model (see Appendix D).

N_THREADS:

Number of concurrent worker threads.

Outputs:

FILE_OUT:

Spreadsheet identical to FILE_IN with an updated column innovation_index_gemini_flash containing a numeric innovation score in [0, 100] for each company, or the string "error".

```
1. LOAD_SPREADSHEET:
     df ← read (FILE_IN)

2. DEFINE FUNCTION BUILD_PROMPT(KEYWORDS):
     PROMPT_INSTANCE:=
       BASE_PROMPT
       + "\n\n"
       + "Company keyword evidence:\n"
       + KEYWORDS
       + "\n\n"
       + "Return ONLY valid JSON with the field"
       + "\"innovation_index\"."
     return PROMPT_INSTANCE

3. DEFINE FUNCTION CALL_LLM_AND_PARSE(PROMPT_INSTANCE):
     RAW_RESPONSE_JSON: = CALL_LLM (model_name      = "innovation-rating-model",
         input_text       = PROMPT_INSTANCE,
         response_schema  = {"innovation_index": float in [0,100]},
         temperature      = 0)
     SCORE: = RAW_RESPONSE_JSON["innovation_index"]
     return SCORE
```

```
4. DEFINE FUNCTION PROCESS_ROW (i, row):
      current_val: = row[“innovation_index“]
      if current_val is not NULL
        and current_val! = “error”:
          return (i, SKIP)

      KEYWORDS: = row[“matched_keywords_string“]
      if KEYWORDS is NULL OR KEYWORDS is “ OR KEYWORDS is “nan”:
          return (i, 0)

      try:
          P: = BUILD_PROMPT(KEYWORDS)
          SCORE:= CALL_LLM_AND_PARSE(P)
          return (i, SCORE)

      catch Exception e:
          return (i, "error")

5. CREATE THREAD_POOL with N_THREADS workers.

6. SUBMIT TASKS:
      FUTURE_MAP:= {}
      for each row index i in df:
          FUTURE:= THREAD_POOL.submit(PROCESS_ROW, i, df[i])
          FUTURE_MAP[FUTURE]:= i

7. INITIALIZE:
      COUNTER:= 0

8. FOR EACH FUTURE as it completes (in completion order):
      i:= FUTURE_MAP[FUTURE]
      COUNTER:= COUNTER + 1

      try:
          (row_idx, result):= FUTURE.result()

          if result == SKIP:
              continue
          else:
              df[row_idx][“innovation_index_gemini_flash”]:= result

      catch Exception e:
              df[i][“innovation_index_gemini_flash”]:= “error”

      if COUNTER mod 1000 == 0:
          SAVE_TO_EXCEL (
            df,
```

```
        filename = "tmp_keywords_" + str(COUNTER) + ".xlsx"
    )

9. AFTER ALL FUTURES COMPLETE:
    SAVE_TO_EXCEL(df, "keywords_final.xlsx")

10. END.
```

D. Prompt specification for innovation scoring

[SECTION A: ROLE / EXPERTISE]

You are an independent innovation assessment expert. Your job is to evaluate how innovative a company appears to be based ONLY on the provided keywords that summarize its public website content (technologies, capabilities, products, R&D focus, patents, etc.).

[SECTION B: WHAT "INNOVATION" MEANS HERE]

When you evaluate "innovation", consider signals such as:

- Novel technology development, R&D activity, patents.
- Use of advanced methods (e.g., AI, automation, proprietary tooling).
- Creation of new products / platforms, not just reselling others'.
- Evidence of continuous improvement, experimentation, prototyping, or scientific / engineering work.
- Unique domain expertise or specialized IP.

Lower innovation means generic services, commodity reselling, or purely operational/administrative wording without signs of original technical problem solving.

[SECTION C: HOW TO SCORE]

You must output a single numeric score called "innovation_index". This score MUST be a float between 0 and 100 (inclusive):

- 0 → almost no evidence of innovation. Mostly generic / routine / legacy activities.
- 20 → minor innovation signals (some customization, limited technical differentiation).
- 50 → moderate innovation signals (clear technical work, product development, R&D-like language).
- 80 → strong innovation focus (specialized technology, advanced methods, in-house development, patents).
- 100 → extremely innovation-driven (cutting-edge research, proprietary breakthroughs, repeatable R&D culture).

Do NOT normalize relative to other companies in the dataset. Score each company independently and absolutely.

[SECTION D: OUTPUT FORMAT (STRICT)]

Your entire response MUST be valid JSON with EXACTLY one key:

{

"innovation_index": <float between 0 and 100>

}

No explanations. No additional fields. No text before or after.
If you are uncertain, make your best estimate.

[SECTION E: SAFETY / MISSING DATA]
If the keywords are missing, empty, or contain no usable technical/business content, respond with:
{
“innovation_index”: 0
}
[SECTION F: COMPANY KEYWORD EVIDENCE]
Company keyword evidence:
<INSERT matched_keywords_string HERE>