The Hidden Innovators: Uncovering Capabilities in European Rural Areas

Online job vacancy data for innovation and industrial dynamics

Bernardo Caldarola[1] Tommaso Ciarli[2] Simone Sasso[1]
Šimon Trlifaj[2] Emanuele Pugliese[2] Alessio Bumbea[2,3]

[1]European Commission
[2]United Nations University-MERIT, Maastricht University
[3]Universitas Mercatorum

All findings preliminary
Thessaloniki, 28 May 2026

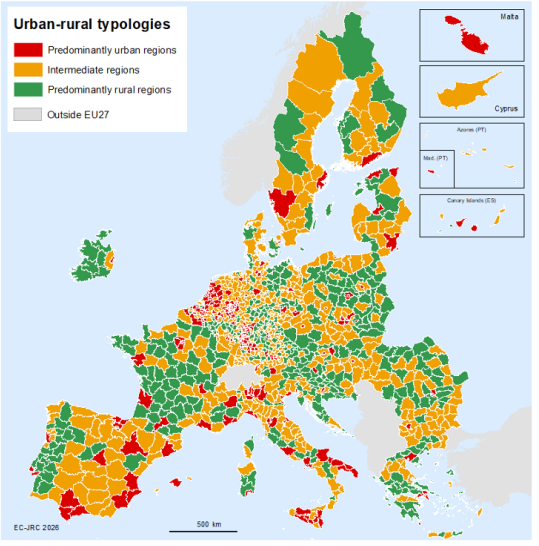
Motivation

EU rural regions account for approx. 45 % of EU land and 21 % if population.

- Increasingly challenging demographic trends and **urban-rural gaps**

⇕

geographies of discontent



Motivation

Innovation in rural and peripheral regions is difficult to observe: standard indicators are often poorly suited to capture innovation dynamics outside major urban centres.

- We use OJV-based skill demand to identify innovation capabilities and uncover 'hidden innovators' across European rural regions.
- OJV data are timely, granular, and directly linked to the knowledge and capabilities firms seek to build.

Hypothesis: hiring for specific skills can provide a leading high-definition signal of innovative activity, even in the absence of patenting.

Goal

Can online job vacancy (OJV) data serve as a signal of innovation?

Innovation is hard to measure on a granular level, especially in remote areas. We use OJV skills to measure firms' revealed skill demand, link these firms to ORBIS, ORBIS IP and PATSTAT, and test whether firm-level skill-demand predict (technology-specific) patenting.

- **Example:** in Regen in Bavaria (Germany), there were 391 patents applied for between 2014 and 2020, and 10,686 OJVs posted in our data.

Background and Literature

- **Geography of innovation and rural peripheries:** innovation is spatially uneven, but peripheral regions still innovate through place-specific capabilities and linkages (Eder, 2019; Andrews and Whalley, 2022).
- **Smart specialization and relatedness:** regional development depends on building on existing capabilities and moving into related activities (Foray et al., 2011; Foray, 2014; Balland et al., 2019; Hidalgo and Hausmann, 2009).
- **Skills as capability signals:** occupations and skills reveal knowledge structures relevant for technological change and firm diversification (Neffke and Henning, 2013; Dotzel and Wojan, 2022; Wirkierman et al., 2024; Mann and Loveridge, 2022).

Contribution: use OJV-based skill demand as a scalable measure of firm capabilities and test whether it predicts subsequent patent outcomes across countries.

Approach

We want to link firms' skill demand to changes in innovative behavior.

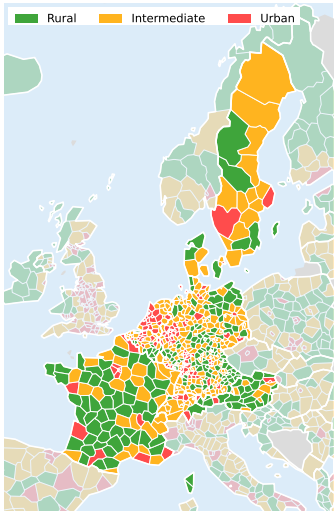
1. Use Lightcast vacancies to observe firm-level skills demand.
2. Through ORBIS (IP), link skill-demand profiles to patenting.
3. Identify skills combinations associated with later innovation.
4. Extrapolate: measure innovative behavior of firms that did not patent.

Approach: patents give us an observed innovation outcome for some firms; OJV data may let us learn the underlying capability profile that precedes it for many more firms.

Data: Lightcast OJV

Skills demand from Lightcast:

- 10 million OJVs from nine European countries[a], 2014–2017
- Median of 32 thousand skills mentions per year per region (16 thousand in rural regions)
- Skills: Lightcast skills prefered labels (ESCO-based, 1,741 values), mapped to ESCO l2 skills (125 values)[b]
- We exclude OJVs by 'employment activities' firms.



[a] AT, BE, CH, DE, DK, FR, LU, NL, SE

[b] Badort A., Caldarola B., Ciarli T., Rony S., (2026), 'Skills Demand and Regional Productivity', TPI deliverable

Data: Agencies in Lightcast OJV

Table: Proportion of OJVs by employment agencies

Country	Vacancies	Share of agencies
NL	10,388,605	66.8%
BE	11,959,555	66.2%
FR	34,751,205	57.5%
CH	9,064,650	46.6%
LU	676,526	40.2%
DE	58,314,588	40.2%
SE	5,654,597	35.3%
AT	3,801,667	31.2%
NO	944,915	25.5%
DK	3,139,230	9.3%

Employment agencies generate a large share of postings in some countries.

Data processing: Matching Lightcast to ORBIS

Lightcast ❶ ORBIS ❷ ORBIS IP ❸ PATSTAT

ORBIS and ORBIS IP provide firm and patenting information.

- We match about half of firms in Lightcast to ORBIS.
- We require NUTS 2-level match between the OJV and at least one of the patent applicants or inventors.
- Matching is balanced in patents WIPO technological classes.

Data processing: Matching Lightcast to ORBIS

Table: Firm-level match rate

Country	Lightcast companies excl. agencies	Matched to Orbis	Match rate Orbis
DE	2,695,861	1,418,383	52.6%
FR	931,933	439,701	47.2%
BE	366,777	175,371	47.8%
CH	310,102	163,157	52.6%
NL	204,535	112,485	55.0%
AT	241,504	131,274	54.4%
SE	205,824	115,630	56.2%
DK	196,962	142,419	72.3%
LU	19,394	9,482	48.9%

Data processing: Matching Lightcast to ORBIS

The matching rate is relatively stable in rural regions.

Table: Firm-level match rate, employers in rural regions

Country	Lightcast companies excl. agencies	Matched to Orbis	Match rate Orbis
DE	205,745	106,998	52.0%
FR	248,739	98,927	39.8%
BE	10,198	5,064	49.7%
CH	21,884	7,314	33.4%
NL	1,519	692	45.6%
AT	51,471	31,513	61.2%
SE	18,212	9,390	51.6%
DK	23,368	17,250	73.8%
LU	0	0	nan%

Descriptives

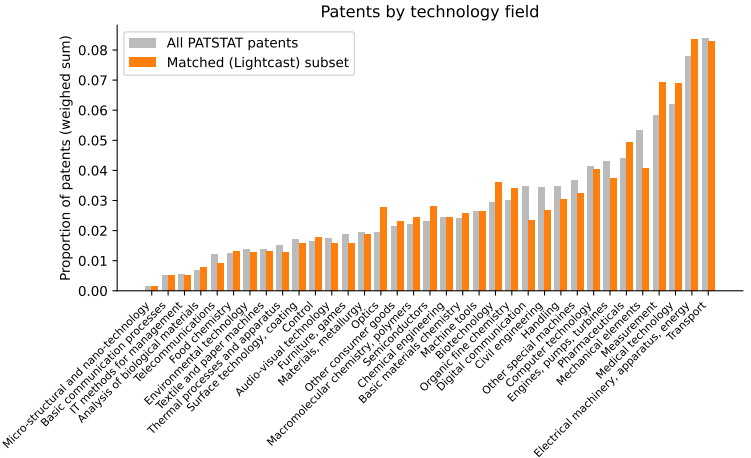
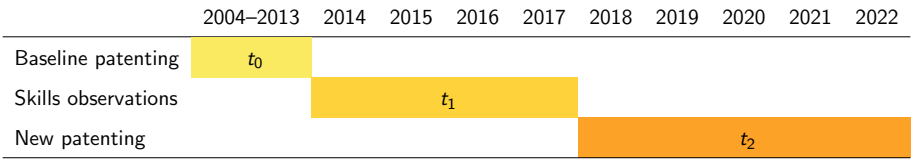


Figure: Distribution of WIPO technology classes between PATSTAT patents and Lightcast matched subsample.

Problem definition: timing

We want to predict if observing jobs demand of firm i in t_1 is predictive of patenting in t_2, treating t_0 as the baseline.



- non-overlapping time windows
- skills observations aggregated over four years in t_1
- firm-level outcomes for patent entry and field entry in t_2

Problem definition: outcomes

Simple extensive margin outcome:

- Patent entry **on a subset of firms that did not patent in** t_0

$$\text{patent\_entry}_{it_2} = 1\{\text{firm } i \text{ patented in } t_2\}$$

Directional extensive margin in two steps:

- Patenting in t_2 (all firms)

$$\text{patenting}_{it_2} = 1\{\text{firm } i \text{ patented in } t_2\}$$

- WIPO technological field entry t_2 **on a subset of firms that did patent in** t_2, but not in field w in t_0

$$\text{tech\_entry}_{iwt_2} = 1\{\text{firm } i \text{ patented in tech. } w \text{ in } t_2\}$$

Problem definition: skills predictors

- For all outcomes, we use skills-based predictors:

$$\text{skills}_{i,k} = \frac{1}{n_i} \sum_{v \in \mathcal{V}_i} \mathbf{1}\{\text{vacancy } v \text{ contains skill } k\}$$
$$\text{skills\_scale}_i = \log(1 + n_i)$$

- For outcomes patenting_{it_2} and $\text{tech\_entry}_{iwt_2}$, we add baseline patenting:

$$\text{patent\_scale}_{it_0} = \log(1 + Npat_{it_0})$$
$$\text{patent\_scope}_{it_0} = \log(1 + Nwto_{it_0})$$

Where $\mathcal{V}_{it}$ is the set of vacancies posted by firm i in t_1 and $n_i t$ is the number of vacancies. $Npat_{it_0}$ and $Nwto_{it_0}$ is the number of patents and distinct WIPO technological fields in t_0.

Descriptives

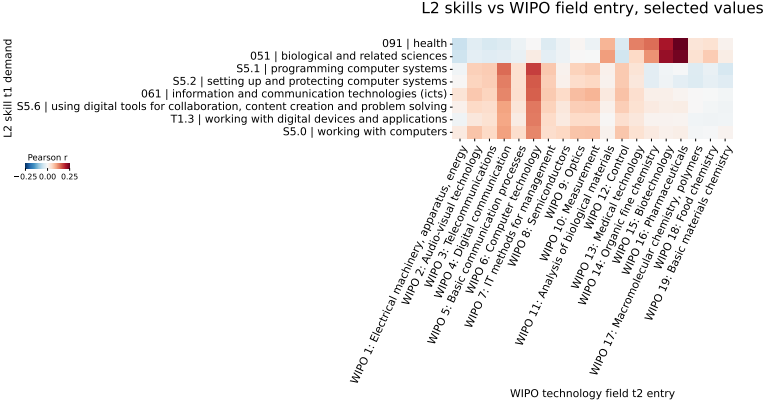


Figure: Correlation between skills demand in t_1 and WIPO technology field entry in t_2, selected categories.

Method: model

We train gradient-boosted decision trees to predict $\text{patent\_entry}_{it_2}$, patenting_{it_2} and $\text{tech\_entry}_{iwt_2}$ from $\text{skills}_{i,k}$, skills_scale_i, $\text{patent\_scale}_{it_0}$, and $\text{patent\_scope}_{it_0}$. Features:

- Inverse-prevalence weighting of events
- Prevalence-aware metrics such as AUPRC and lift
- Held-out performance across countries for patent-entry prediction
- Comparison to random models
- SHAP values for features and their combinations

Results: Patenting entry in t_2

We predict patent_entry$_{it_2}$ only from skills demand in t_1: skills$_{i,k}$, skills_scale$_i$. This is a rare outcome: only 0.46 % of firms enter patenting in t_2.

Table: Model performance: patenting entry

Model	AUPRC	AUPRC / prev.	Lift 1%	ROC AUC
LightGBM	0.016	4.532	7.560	0.707
Random baseline	0.004	1.043	1.238	0.502

The model is about 4.5 times better than random baseline.

Results: t_2 patenting

We predict patenting$_{it_2}$ from skills demand and prior patenting. This is still a rare outcome: 1.1 % of firms enter patenting in t_2.

Table: Model performance: patenting in t_2

Model	AUPRC	AUPRC / prev.	Lift 1%	ROC AUC
LightGBM	0.476	70.075	51.820	0.886
Random baseline	0.011	1.004	0.998	0.500

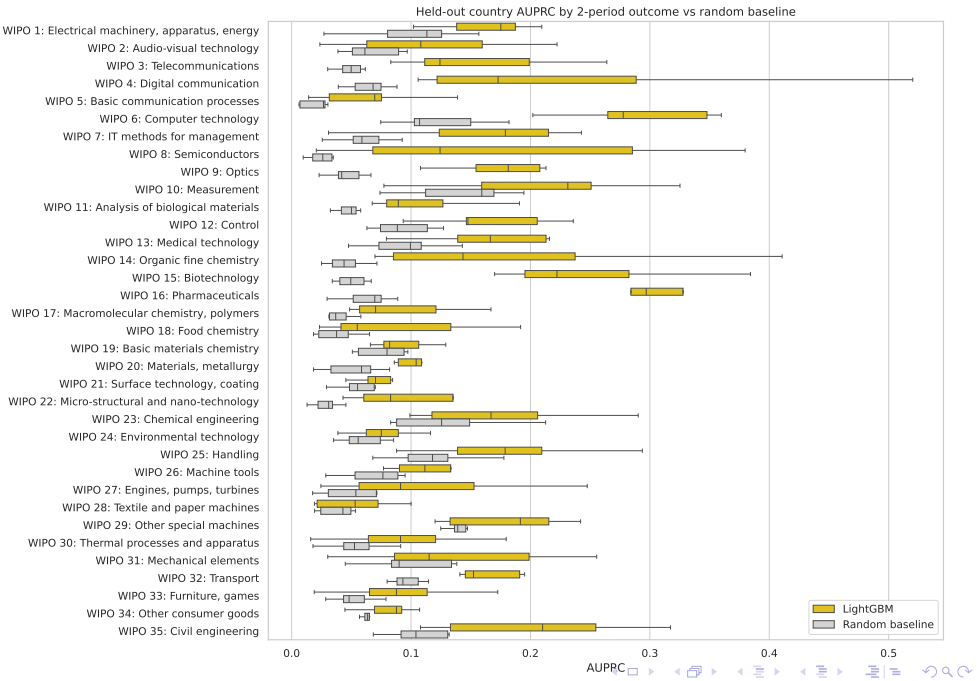
The model is about 70 times better than random baseline.

Results: WIPO field entry in t_2 | t_2 patenting

We predict tech_entry$_{iwt_2}$ for WIPO technology fields, from skills demand and prior patenting. The models are on average 2.7 times better than random (max 7.6, min 1.3).
Example:

Table: Model performance: WIPO 15 Biotechnology entry in t_2 | t_2 patenting

Model	AUPRC	AUPRC / prev.	Lift 1%	ROC AUC
LightGBM	0.274	6.057	13.721	0.720
Random baseline	0.050	1.028	0.157	0.498


Held-out country AUPRC by 2-period outcome vs random baseline
WIPO 1: Electrical machinery, apparatus, energy
WIPO 2: Audio-visual technology
WIPO 3: Telecommunications
WIPO 4: Digital communication
WIPO 5: Basic communication processes
WIPO 6: Computer technology
WIPO 7: IT methods for management
WIPO 8: Semiconductors
WIPO 9: Optics
WIPO 10: Measurement
WIPO 11: Analysis of biological materials
WIPO 12: Control
WIPO 13: Medical technology
WIPO 14: Organic fine chemistry
WIPO 15: Biotechnology
WIPO 16: Pharmaceuticals
WIPO 17: Macromolecular chemistry, polymers
WIPO 18: Food chemistry
WIPO 19: Basic materials chemistry
WIPO 20: Materials, metallurgy
WIPO 21: Surface technology, coating
WIPO 22: Micro-structural and nano-technology
WIPO 23: Chemical engineering
WIPO 24: Environmental technology
WIPO 25: Handling
WIPO 26: Machine tools
WIPO 27: Engines, pumps, turbines
WIPO 28: Textile and paper machines
WIPO 29: Other special machines
WIPO 30: Thermal processes and apparatus
WIPO 31: Mechanical elements
WIPO 32: Transport
WIPO 33: Furniture, games
WIPO 34: Other consumer goods
WIPO 35: Civil engineering
0.0
0.1
0.2
0.3
0.4
0.5
AUPRC
LightGBM
Random baseline

Results: Entry into patenting

We calculate SHAP values for interactions of features in each model, which gives us a ranking of predictors.

Table: Entry into patenting, top 10 SHAP features

Feature 1	Feature 2	Importance
I2: engineering and engineering trades	I2: working with computers	0.0185
I2: working with computers	I2: using digital tools for collaboration, content creation and problem solving	0.0174
I2: engineering and engineering trades	I2: physical sciences	0.0159
I2: working with computers	I2: collaborating in teams and networks	0.0143
I2: providing information and support to the public and clients	I2: using digital tools for collaboration, content creation and problem solving	0.0107
I2: providing information and support to the public and clients	I2: manufacturing and processing	0.0090
I2: engineering and engineering trades	I2: communication, collaboration and creativity	0.0089
I2: engineering and engineering trades	I2: mastering languages	0.0086
I2: working with computers	I2: law	0.0085
I2: working with computers	I2: providing information and support to the public and clients	0.0083

Results: WIPO 6 features interpretation

Table: WIPO 6 Computer technology, top 10 SHAP features

Feature 1	Feature 2	Importance
l2: information and communication technologies (icts)	l2: setting up and protecting computer systems	0.0233
patent baseline: log1p_n_t0	patent baseline: log1p_n_fields_t0	0.0191
l2: information and communication technologies (icts)	patent baseline: any_t0	0.0123
patent baseline: log1p_n_t0	l2: setting up and protecting computer systems	0.0121
patent baseline: log1p_n_t0	l2: engineering and engineering trades	0.0119
l2: setting up and protecting computer systems	l2: using digital tools for collaboration, content creation and problem solving	0.0105
l2: programming computer systems	patent baseline: any_t0	0.0104
patent baseline: log1p_n_t0	l2: information and communication technologies (icts)	0.0100
patent baseline: log1p_n_fields_t0	l2: information and communication technologies (icts)	0.0096
patent baseline: log1p_n_t0	l2: demonstrating willingness to learn	0.0093

Results: Extrapolating innovation capabilities

Do these models give us new information about innovation in rural regions?

We compare patent-based and skills-based innovative activity across NUTS3 regions by WIPO technology fields.

- *Patent-based* activity: at least 5 firms in the NUTS3 patented in WIPO field w in 2019.
- *Skills-based* activity: at least 5 firms in the NUTS3 are above the 95th percentile of the predicted score for entry into a WIPO field w.

Extrapolation: The skills-based score is obtained from the WIPO-field entry model tech_entry_{iwt}. We use all matched firms with observed skill demand in the selected year, including firms with no patenting. Percentile across regions, 2019 data.

Results: Coverage urban

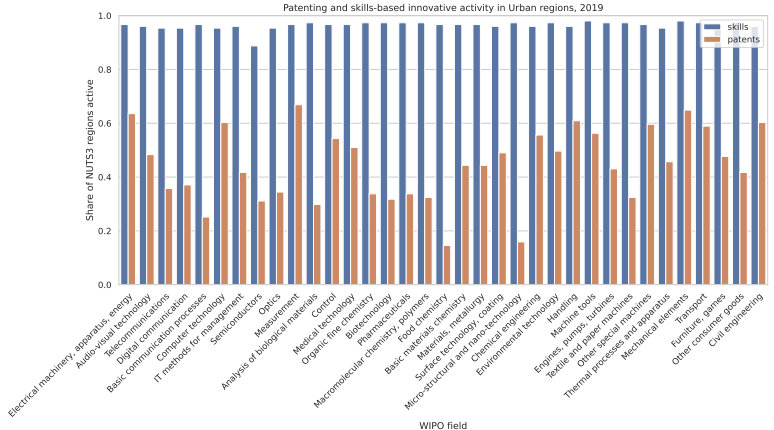


Figure: Patents- and skills-based innovative activity, urban regions.

Results: Coverage intermediate

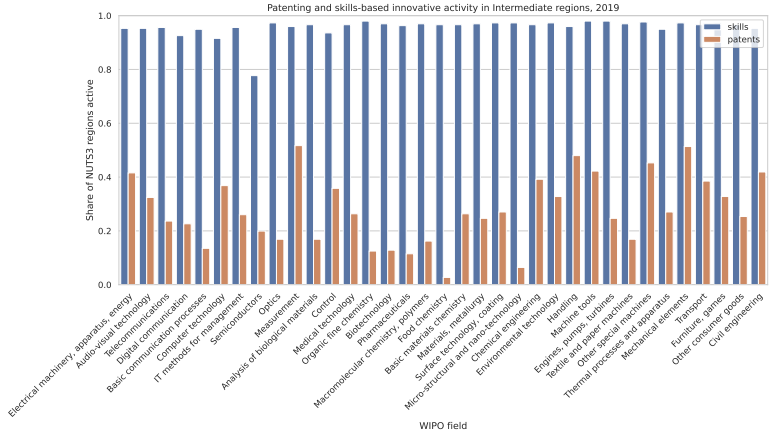


Figure: Patents- and skills-based innovative activity, urban regions.

Results: Coverage rural

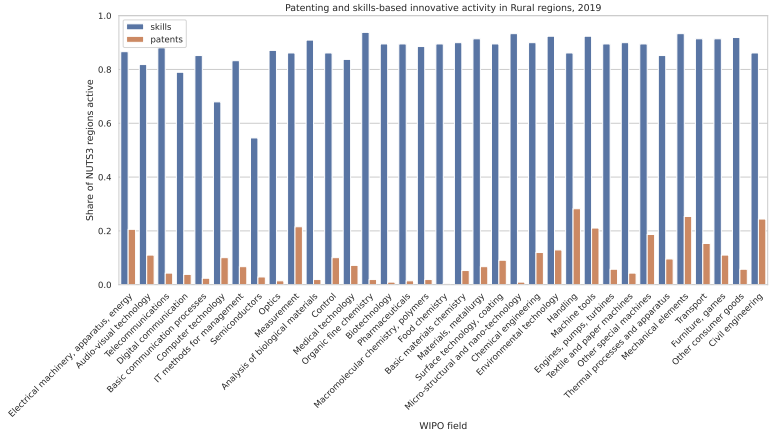


Figure: Patents- and skills-based innovative activity, urban regions.

Results: innovation-related skills by regional typology

Do urban, intermediate and urban regions show different representation of innovative skills?

We take the top 10 skills features from the patent entry model and WIPO tech class entry models, and plot their relative representation by territorial type.

Results: Skills by regional typology, patent entry skills

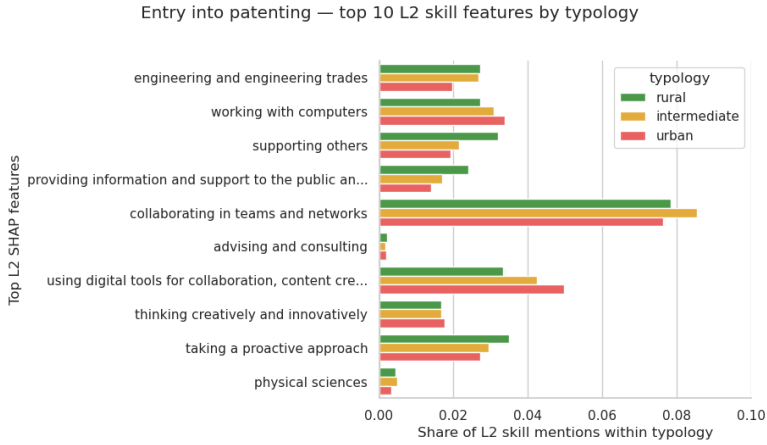


Figure: Innovative skills

Results: Skills by regional typology, entry in computer technology

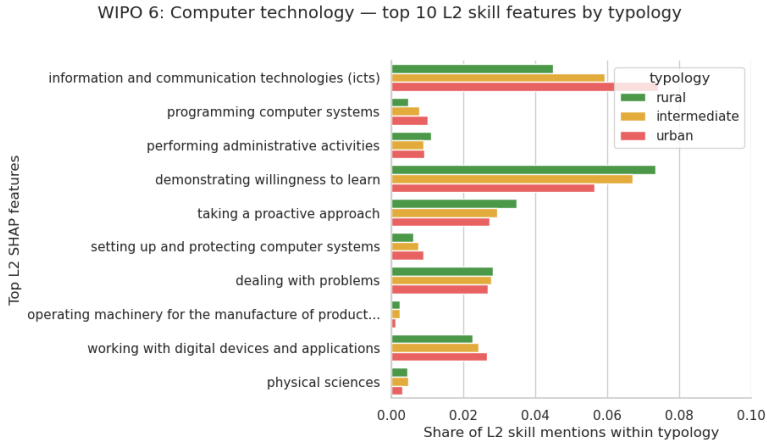


Figure: WIPO 6 innovative skills

Takeaways

1. **Online job vacancies provide a high-resolution lens** on firms' innovation capabilities and capability-building processes.

2. Firms' skill demand contains meaningful predictive information about **future patenting** and **technological diversification**.

3. Innovation capabilities appear much more **territorially distributed** than patent data alone would suggest, especially in rural regions.

4. OJV data can help identify **"hidden innovators"** and support more **place-sensitive innovation policy**.

References I

Andrews, M. J. and A. Whalley (2022, May). 150 years of the geography of innovation. *Regional Science and Urban Economics 94*, 103627.

Balland, P.-A., R. Boschma, J. Crespo, and D. L. Rigby (2019, September). Smart specialization policy in the European Union: relatedness, knowledge complexity and regional diversification. *Regional Studies 53*(9), 1252–1268.

Dotzel, K. R. and T. R. Wojan (2022). An Occupational Approach to Analyzing Regional Invention. Technical Report NCSES 22-202, National Science Foundation, National Center for Science and Engineering Statistics (NCSES).

Eder, J. (2019, March). Innovation in the Periphery: A Critical Survey and Research Agenda. *International Regional Science Review 42*(2), 119–146.

Foray, D. (2014, January). From smart specialisation to smart specialisation policy. *European Journal of Innovation Management 17*(4), 492–507.

Foray, D., P. David, and B. Hall (2011). Smart specialization: From academic idea to political instrument, the surprising career of a concept and the difficulties involved in its implementation. *MTEI Working Paper*.

Hidalgo, C. A. and R. Hausmann (2009, June). The building blocks of economic complexity. *Proceedings of the National Academy of Sciences 106*(26), 10570–10575.

References II

Mann, J. and S. Loveridge (2022, October). Measuring urban and rural establishment innovation in the United States. *Economics of Innovation and New Technology 31*(7), 650–667.

Neffke, F. and M. Henning (2013, March). Skill relatedness and firm diversification. *Strategic Management Journal 34*(3), 297–316.

Wirkierman, A. L., T. Ciarli, and M. Savona (2024). Employment imbalances in EU regions: technological dependence or high-tech trade centrality? *Regional Studies 59*(1), 2392794.

Appendix

Matching: overview

1 matching:

1. Clean and normalize company names in both data sources.
2. Generate name candidates using embedding similarity, fuzzy matching, and acronym rules.
3. Accept high-confidence deterministic matches using name and geography information.
4. Resolve remaining ambiguous cases with structured LLM-based matching, based on: name, sector, location.

2 through BvD IDs. 3 Through patent family ID.

- We require NUTS 2-level match between the OJV and at least one of the patent applicants or inventors.

Matching: Pre-processing

We retrieve company name and OJV location/sector from Lightcast, and company name, location, and sector from ORBIS.

1. Clean names (suffixes, case, special characters).
2. Aggregate Lightcast by name and NUTS-3. Give each company a unique source ID.
3. Map ZIP codes from ORBIS to NUTS-3 and aggregate on BvD, name, and NUTS-3. Give each company a unique target ID.

- Use all three names (current, previous, alternative) available in ORBIS.
- If multiple BvD IDs exist with the same location and name, attempt to find a common main BvD ID.

Matching: Candidates generation

We generate candidate matches using three methods:

1. **Semantic**: Embed each name using sentence transformer (all-MiniLM-L6-v2). Take top 3 closest names if similarity > 0.8.
   - Example: 'IBM' and 'IBM international' score 0.88.
2. **Fuzzy**: Calculate Jaro-Winkler normalised similarity. Take top 3 if similarity > 0.875.
   - Example: 'Google' and 'Googl' score 0.9.
3. **Acronym**: If a name has more than three words, include all names with a matching abbreviation.
   - Example: 'IBM' and 'International Business Machines'.

Matching: Deterministic & LLM-based match

Deterministic match

1. If name and NUTS-3 produces exactly one match.
2. If semantic similarity > 0.9 + NUTS-3 match (or >0.95 and NUTS-3 unavailable) + similarity to next candidate is at least 0.05.

LLM-based match

- Unresolved companies are submitted to an LLM (GPT 5 nano, low reasoning).
- Requests include source company details and 3 candidate variants.
- Keep matches with confidence > 0.8.