Uncovering Hidden Innovators: Linking administrative and big data to develop comprehensive measures of firms' innovation

Enrico Vanino
School of Economics
University of Sheffield

Laura Haddock
Survey and Economic Indicators Division
ONS

ESCOE 2025 - KCL 23rd May 2025

Introduction

Limitations of Current Innovation Metrics

- **Limitations of current official metrics** to provide clear picture of innovative activities of firms:
  1. **Measures shortcomings**: Focus on narrow definition of innovation (e.g. R&D investment, Frascati definition, patents, etc.);
  2. **Limited scope**: Limited and non-representative samples of firms (e.g. surveys, large firms, HQs), with extensive heterogeneity hidden behind aggregate statistics;
- Filling these 2 gaps crucial to **strengthen data and analytical capacity** to provide comprehensive evidence-based policy suggestions.

Introduction

Measures Shortcomings

- **Shortcomings** of standard **measures of firms innovation**, like R&D investment, employed researchers, patents, etc. (NESTA, 2007);
  - Traditional metrics based on specific model of **science-based innovation** led by formal R&D that is increasingly less relevant;
  - Particularly inappropriate for economies with lower reliance on manufacturing and **greater importance of services**;
- **Majority of innovations** are not based on the latest scientific or technological knowledge, but on exploiting the existing one (Bender and Laestadius, 2005; Arundel et al., 2008).
- Many sectors are more dependent on "**hidden or latent innovation**" than on traditional science-based innovation (Barge-Gil et al., 2011; Goetz and Han, 2020):
  - Incremental innovation;
  - Design and trademarks;
  - Adoption of new technologies and processes;
  - Intra-value chain collaboration;
  - Employees training;

Introduction

Limited Scope

- Limited focus on certain types of innovative activities (e.g. R&D investment and patents) leads to focus on a **narrow (biased?) set of businesses**:
  - **Large firms** operating in manufacturing and/or **high-tech** industries, clustered in densely populated **urban areas**;
- This is also due to **data collection constraints**:
  - Limited information on innovation from balance-sheet data;
  - Need to rely on surveys, with all related limitations:
    - Focus on easily collectable variables;
    - Prone to measurement error;
    - Limited samples representative only at very aggregated level (e.g. broad regional OR industrial classifications);
    - Cross-sectional data or with attrition issues;
- **New alternative unstructured data** sources which could **better represent** businesses innovation:
  - Limited, partial, or sector specific coverage (e.g. GitHub, crowdfunding, web-scraping, etc.)
  - Used in isolation - difficult data linking;
  - Noisy data prone to measurement error.

Aims

Project Objectives

- Develop **new comprehensive measures** of business R&D and innovation activities to identify **hidden innovators** not captured by traditional survey data:
  1. **Different types** of innovative activities (beyond R&D investment & patents);
  2. **Broader coverage** including entire business population (e.g. also SMEs, low-tech industries, peripheral/rural areas);
- **Map**, **gather**, and **link** official **survey** statistics with **administrative** datasets and other **alternative** data sources to create most extensive micro-level database on firms innovative activities.

Benefits

Project Contribution

1. Map **data sources** available, **variables definition** and **comparability** to analyse business innovative activities.
2. Create new **comprehensive database** for business innovation and R&D activities in the UK (and blueprint for other OECD countries).
3. Shed a light on innovation activity of hidden innovators, in line with national statistical offices' initiatives of **transforming R&D and innovation statistics**.
4. Enhance **sampling strategy** and lowering costs for future R&D and innovation surveys.
5. **Strengthen data and analytical capacity** for academic and policy analysis of business innovation.
6. Provide evidence to develop effective **innovation policies** to support businesses.
7. Encourage **engagement and cooperation** between statistical office and government departments enabling future data sharing (e.g. ONS, HMRC, UKRI, Devolved Nations, etc.).

Data

Data on Firms Innovation

- Link longitudinal micro-level data on a broad range of **different businesses R&D and innovation activities**;
- What type of data sources on business R&D and innovation activities can we consider?
  1. **Business R&D Official Surveys** [SRV];
  2. **Administrative Data on Public R&D Support and IPRs** [ADM];
  3. **Alternative Data Sources on Businesses' Innovative Activities** [ALT].

Data

Data Sources UK

Sources of micro-level data on **different UK businesses R&D and innovation activities** between 2015-2020:

1. **Official Surveys Data** [SRV]:
   - Business Enterprise Research and Development [BRD];
   - UK Innovation Survey [CIS];
2. **Administrative Data** [ADM]:
   - HMRC R&D Tax Credit [HMR];
   - National UKRI & Innovate UK Research Funding & Support [GTR];
   - National Catapults Support for Technological Adoption [CAT];
   - Regional Research Support (SMART Wales, Invest NI, Scottish Enterprise, HIENT, INTERFACE) [REG];
   - IPO Intellectual Property Rights (patents, design, trademarks) [IPO];
3. **Alternative Data Sources** [ALT]:
   - DataCity web-scraped innovation index [DCT].

- Complemented with firms balance-sheet data.

Data Analysis

Data Sources Overlap

Tab.1: Share of innovators identified across surveys, administrative and alternative data sources.

	BRD	CIS	DCT	GTR	HMR	IPO	REG
BRD	100.00%	31.27%	9.26%	28.72%	25.34%	11.78%	26.60%
CIS	15.75%	100.00%	2.61%	10.03%	8.37%	4.84%	12.20%
DCT	1.84%	1.03%	100.00%	3.90%	2.30%	1.69%	2.91%
GTR	15.12%	10.49%	10.32%	100.00%	10.59%	7.69%	22.19%
HMR	42.82%	28.09%	19.50%	34.00%	100.00%	19.56%	40.96%
IPO	19.85%	16.18%	14.33%	24.60%	19.49%	100.00%	28.34%
REG	2.16%	1.97%	1.19%	3.43%	1.97%	1.37%	100.00%

Data Analysis

Data Sources Overlap

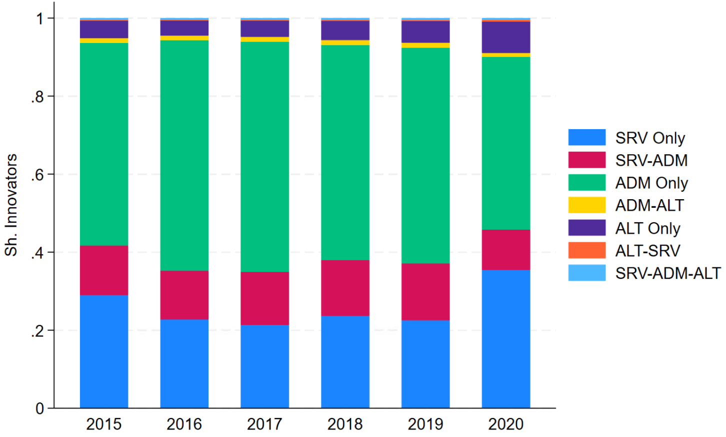
Tab.2: Correlation between innovators definition across surveys, administrative and alternative data sources.

	BRD	**CIS**	**DCT**	**GTR**	**HMR**	**REG**	**IPO**
BRD	1						
CIS	0.1356	1					
DCT	0.038	0.0096	1				
GTR	0.1677	0.0633	0.0496	1			
HMR	0.2634	0.0926	0.0647	0.1642	1		
REG	0.0284	0.0193	0.0087	0.0381	0.0476	1	
IPO	0.0962	0.0473	0.0248	0.092	0.113	0.0223	1

Data Analysis

Data Sources Overlap

Fig.1: Share of total innovators identified across surveys, administrative and alternative data sources.



Data Analysis

Data Sources Characteristics

Tab.3: Comparison between innovating firms characteristics across data sources and in comparison with firms also included in survey data.

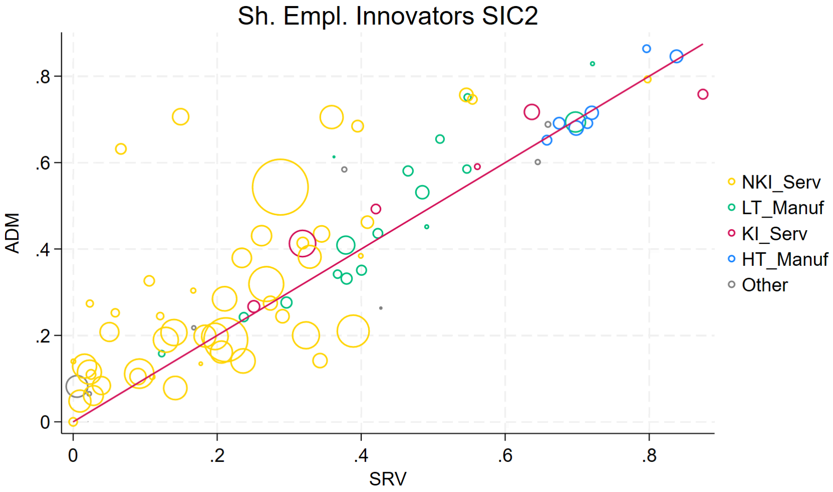
	HMR			IPO			GTR		
	No SRV	**SRV**	**Diff. T-test**	**No SRV**	**SRV**	**Diff. T-test**	**No SRV**	**SRV**	**Diff. T-test**
No. Firms	221,960	67,538		80,982	17,011		48,825	24,848	
Employment	87.20	230.15	15.51	197.00	616.67	13.81	308.01	529.40	6.48
Age	16.65	25.08	160.0	12.03	25.27	130.0	16.17	23.11	66.48
Group	28.33%	58.16%	150.0	27.53%	66.27%	100.0	39.69%	60.96%	55.93
For.Own.	3.79%	13.81%	96.59	5.07%	17.72%	58.79	14.22%	22.65%	28.93
Manuf.	19.74%	45.29%	140.0	12.58%	43.69%	100.0	15.60%	38.41%	71.62
HT Manuf	5.60%	22.28%	130.0	3.54%	21.12%	87.14	7.68%	23.32%	61.30
Service	78.22%	54.29%	125.4	86.68%	55.82%	98.39	81.47%	60.77%	62.54
KIS	18.79%	21.60%	16.17	10.81%	20.03%	33.33	20.63%	32.27%	34.97
Urban	76.95%	78.04%	5.92	81.08%	78.35%	8.17	77.65%	75.91%	5.27

	REG			CIS			DCT		
	No SRV	**SRV**	**Diff. T-test**	**No SRV**	**SRV**	**Diff. T-test**	**No SRV**	**SRV**	**Diff. T-test**
No. Firms	3,751	1,051		3,734	2,981		44,142	4,184	
Employment	65.61	215.07	3.40	460.30	800.20	2.58	16.27	191.83	17.03
Age	16.18	17.45	2.64	15.35	23.55	22.89	10.74	17.31	45.72
Group	24.21%	40.91%	10.80	41.54%	63.07%	17.95	21.79%	47.49%	37.74
For.Own.	3.71%	10.85%	9.24	14.57%	22.31%	8.24	5.94%	12.67%	16.36
Manuf.	36.87%	39.58%	1.61	24.26%	44.15%	17.59	3.39%	11.45%	25.26
HT Manuf	11.14%	20.55%	8.01	10.34%	27.74%	18.90	1.79%	8.29%	26.72
Service	61.50%	58.99%	1.48	75.07%	55.32%	17.39	96.41%	85.88%	24.17
KIS	13.60%	27.02%	10.47	13.69%	32.77%	19.22	45.92%	62.93%	21.15
Urban	65.52%	68.89%	1.63	78.13%	77.28%	0.83	81.86%	82.27%	0.66

Data Analysis

Survey v Admin Data - Industries

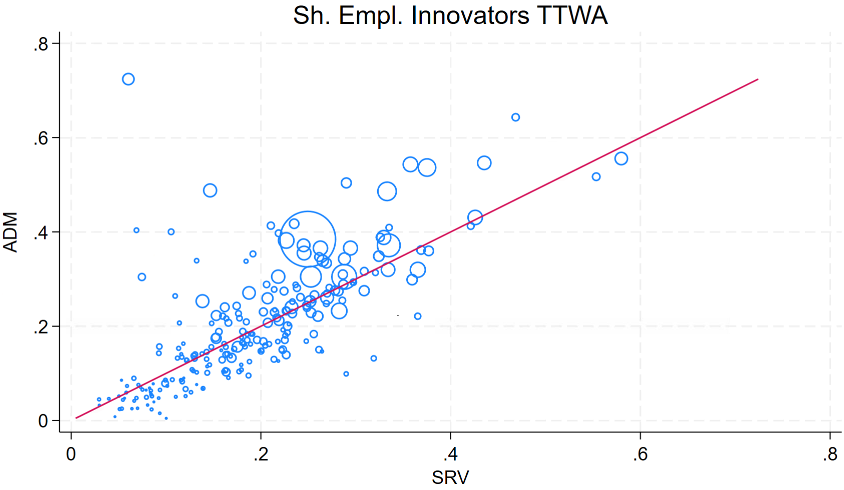
Fig.2: Share of innovators across Industries identified using Survey or Administrative data sources.



Data Analysis

Survey v Admin Data - Regions I

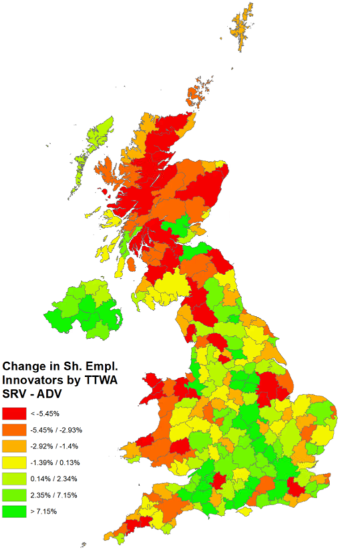
Fig.3: Share of innovators across TTWAs identified using Survey or Administrative data sources.



Data Analysis

Survey v Admin Data - Regions II

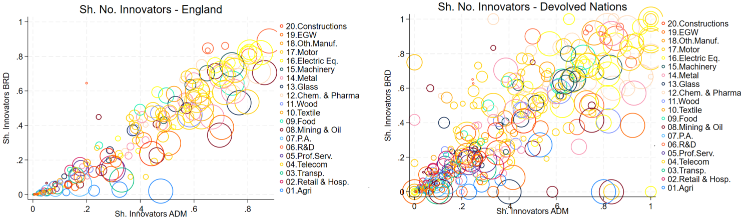
Fig.4: Change in share of innovators across TTWAs identified using Survey or Administrative data sources.



Data Analysis

Survey v Admin Data - Size, Nations & Industries

Fig.5: Share of innovators across industry and firms size groups identified using Surveys or Administrative data sources – England and Devolved Nations.



Data Analysis

Hidden Innovators Characteristics

Tab.4: Regression analysis estimating probability of being an hidden innovator based on firms' characteristics.

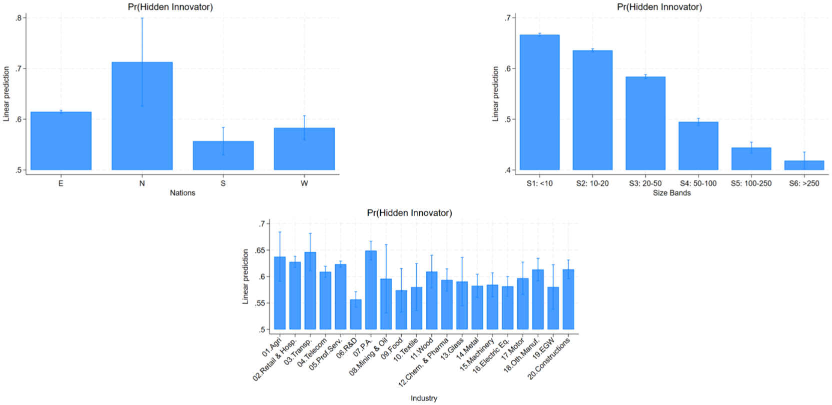
	pr(Hidden Innovator)
Size	-0.0435***
	(-0.00122)
Devolved	-0.0426***
	(-0.00959)
Service	0.0721***
	(-0.0092)
High-Tech	-0.0258***
	(-0.00436)
Firm FE	Y
Year FE	Y
Observations	504,095
R-squared	0.896

Notes: Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$

Data Analysis

Hidden Innovators Characteristics

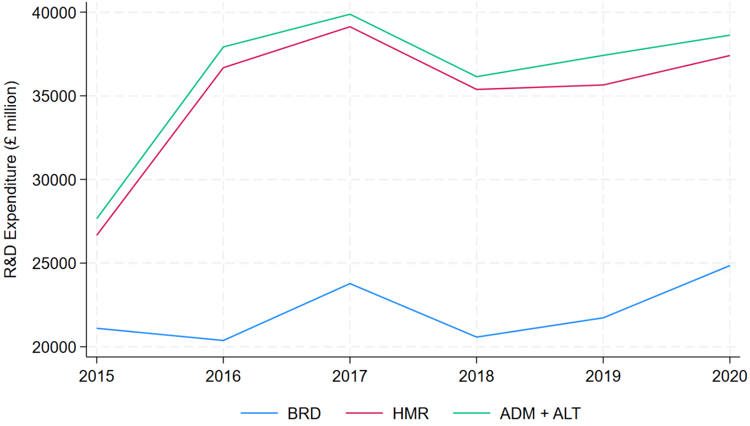
Fig.6: Probability of being an Hidden Innovator by nation, size category, and industry.



Data Analysis

Hidden R&D Expenditure

Fig.7: R&D expenditure estimated using Surveys or Administrative data sources.



Conclusions

Summary & Insights

- Some overlap between SRV and ADM data, mainly through HMR data.
- SRV capture 40% of innovators → 65k new hidden innovators identified using ADM+ALT data.
- Hidden innovators very different from SRV: younger, smaller, single plant, domestic, in low-tech, and services (except for REG).
- ADM data much better at identifying hidden innovators in low-tech services industries and in larger metropolitan areas.
- In England ADM and SRV yield consistent results, very large discrepancies in devolved nations.
- Not because of location, but because fo their characteristics: smaller, services, low-tech firms in England and NI more likely to be hidden innovators.
- SRV data significantly underestimates business R&D expenditure: about £10b more using HMR data, + £1/2b more using otehr ADM+ALT data.
- Value for money in integrating ADM and SRV data for analysis and policymaking.

Thank you

Feedback & Questions?

Contacts: e.vanino@sheffield.ac.uk; Laura.Haddock@ons.gov.uk

Final Report will be published in June 2025

Acknowledgments: This project was supported by the British Academy Innovation Fellowship funding scheme. We thank the following organisations for their collaboration and for sharing their data: ONS, HMRC, UKRI, IPO, the Welsh Government, Invest NI, Scottish Enterprise, INTERFACE, HIENT, and DataCity. We are grateful for the feedback received on previous versions of this work from GEOINNO 2024, DSIT, ONS, AAG 2024, ESCOE 2024, and the EC-JRC. Special thanks to Laura Requena and Beck Keane of the ONS Survey and Economic Indicators Division for their support.