Mapping information economy firms with Big Data: findings for the UK

Max Nathan[*] and Anna Rosso[**]

[*] LSE and NIESR
[**] NIESR

RSA, LONDON, 28 NOVEMBER 2014

Overview

- **Focus:** we look at ICT-producing firms across the UK, aka ‘information economy’ firms
- We generate new info on company counts and characteristics

- **How**: we use a mix of administrative data and ‘big data’, provided by the data science firm Growth Intelligence
- **Findings**: We find 42% more ICT-producing firms than estimates using conventional data, plus UK-wide clusters

- **Why do we care?** This stuff is hard to do. So there’s policy interest (clusters, industrial policy); pros/cons of ‘frontier data’

What does the UK’s ‘information economy’ look like?

SIC07	SECTOR NAME
26	**Manufacture of computer, electronic and optical products**
58	**Publishing activities**
5821	Publishing of computer games
5829	Other software publishing
61	**Telecommunications**
62	**Computer programming, consultancy and related activities**
63	**Information service activities**
6311	Data processing, hosting and related
6312	Web portals

Silicon Valley
electricworks
URBAN AGE
ELECTRIC CITY
CONFERENCE
URBAN AGE
ELECTRIC CITY
CONFERENCE

Could big data help?

- We often need to understand sector / cluster shapes. **It’s a precondition for effective policy design**

- **It’s hard to do this using conventional tools**. SMEs don’t show up in the best UK data (the BSD); other data is missing crucial info (21% of Companies House obs lack SIC codes).
- **More fundamental issue**: real world characteristics of industries evolve faster than coding systems

- Can ‘frontier data’ help here?

The Growth Intel dataset

- **Mix of public and proprietary ‘big data’**

- **Public layer** = Companies House, the UK companies register

- Can be updated daily from the Companies House API feed
- 3.07m raw observations (active companies in August 2012)

- Match to structured data and to modelled data layers

The Growth Intel dataset (2)

- **Structured layer** = matched patents, TMs, exports etc.

- **Unstructured layer** = crawled from internet and newsfeeds

- Matched on company name, address, other observables
- **Text and data mining** => tokens, categories, relevance
- **Supervised learning =>** modelled 'events'; modelled revenue; product / sector / client / platform typologies

Pros and cons

- **Coverage**: 100% of active companies, can be updated daily
- **Granularity**: 5510 sector*product cells vs 806 industry sectors
- Potential to do **text mining** using tokens / event raw data
- **Modelled ‘events’** give us some insight into mergers; joint ventures; product launches; personnel changes and so on

- Need to (extensively) **clean** Companies House data
- **Partial coverage** of employment, revenue, trading addresses
- Coverage of **web-scraped data**

Build

- **Building a benchmarking sample – compare SIC and Gi-based estimates of information economy firms**

- Include only obs with SIC and GI info => smaller than ‘true’
- Benchmarking sample = 1.94m ‘quasi-enterprises’

- Robustness checks with alternative selection techniques

- Validate ‘true’ sample (2.254m) levels and internal structure against BSD and BPE administrative databases

Mapping exercise

- **Identifying ‘ICT producing firms’**
- Compare estimates using (self-assessed) SICs, vs modelled Gi sector and product information

1) Start with companies with information economy SIC codes
2) Extract corresponding Gi sector and product codes
3) Exclude ‘sparse’ sectors/products, recover ‘relevant’
4) Create sector*product cells = **companies in ‘info economy sectors’ whose *principal output* is an information economy good/service**

Company counts

	Observations	%
A. SIC 07 - manufacturing and services		
Other	1,783,973	91.83
Information Economy	158,810	8.17
B. Gi sector and product - manufacturing and services		
Other	1,716,983	88.38
Information Economy	225,800	11.62
Total	1,942,783	100

Robustness checks

- We then run various **robustness checks**:

- Vary the starting set of SIC codes (very narrow, very broad) – we don’t want results to be driven by this
- Re-run the exercise with just sector / just product cells – estimates should blow up
- Vary our exclusion rules (make them tighter, so exclude more)
- Use text-mining to look at keywords in the largest GI cells

Sector breakdowns

	Observations	*%*
information_technology	104,768	46.4
mechanical_or_industrial_engineering	27,326	12.1
computer_software	23,455	10.39
electrical_electronic_manufacturing	17,319	7.67
telecommunications	15,237	6.75
marketing_advertising	11,038	4.89
design	10,049	4.45
computer_networking	3,902	1.73
computer_hardware	3,514	1.56
internet	2,954	1.31
computer_games	2,585	1.14
consumer_electronics	2,074	0.92
information_services	823	0.36
e_learning	347	0.15
computer_network_security	226	0.1
semiconductors	183	0.08
Total	*225,800*	*100*

Product breakdowns

	Observations	***%***
consultancy	151,408	67.05
custom_software_development	19,981	8.85
care_or_maintenance	15,663	6.94
electronics	15,180	6.72
broadband_services	8,628	3.82
web_hosting	6,021	2.67
software_desktop_or_server	5,237	2.32
advertising_network	1,663	0.74
peer_to_peer_communications	1,300	0.58
education_courses	645	0.29
software_web_application	43	0.02
software_mobile_application	31	0.01
Total	*225,800*	*100*

IE companies’ revenue growth in 2010-2012 is faster than non-IE ...

	A. Average Revenues		B. Revenue growth / year	
	Companies House		Companies House (%)	
	mean	median	mean	median
SIC 07				
Other	19,140,919	116,067	0.16	0.02
ICT MF and services	9,760,607	95,400	0.23	0.05
GI sector and product				
Other	19,083,211	115,271	0.16	0.02
ICT MF and services	13,303,007	102,551	0.22	0.05

IE companies’ revenue growth in 2010-2012 is faster than non-IE ...

	A. Average Revenues		B. Revenue growth / year	
	Companies House		Companies House (%)	
	mean	median	mean	median
SIC 07				
Other	19,140,919	116,067	0.16	0.02
ICT MF and services	9,760,607	95,400	0.23	0.05
GI sector and product				
Other	19,083,211	115,271	0.16	0.02
ICT MF and services	13,303,007	102,551	0.22	0.05

IE companies' revenue growth in 2010-2012 is faster than non-IE...

	A. Average Revenues		B. Revenue growth / year	
	Companies House		Companies House (%)	
	mean	median	mean	median
SIC 07				
Other	19,140,919	116,067	0.16	0.02
ICT MF and services	9,760,607	95,400	0.23	0.05
GI sector and product				
Other	19,083,211	115,271	0.16	0.02
ICT MF and services	13,303,007	102,551	0.22	0.05

... but employment data is a bit more mixed.

	2010-2012			
	Obs	**Mean**	**Median**	**% all jobs**
A. SIC07				
Other	70,805	21.13	4	96.23
ICT mf and services		17.06	2	3.77
B. GI sector/product				
Other	70,805	20.88	4	92.86
ICT mf and services		21.65	3	7.14

Note: ***sub-sample of firms reporting employment to Companies House.*** *Data is averaged over 2010-2012.*

... but employment data is a bit more mixed.

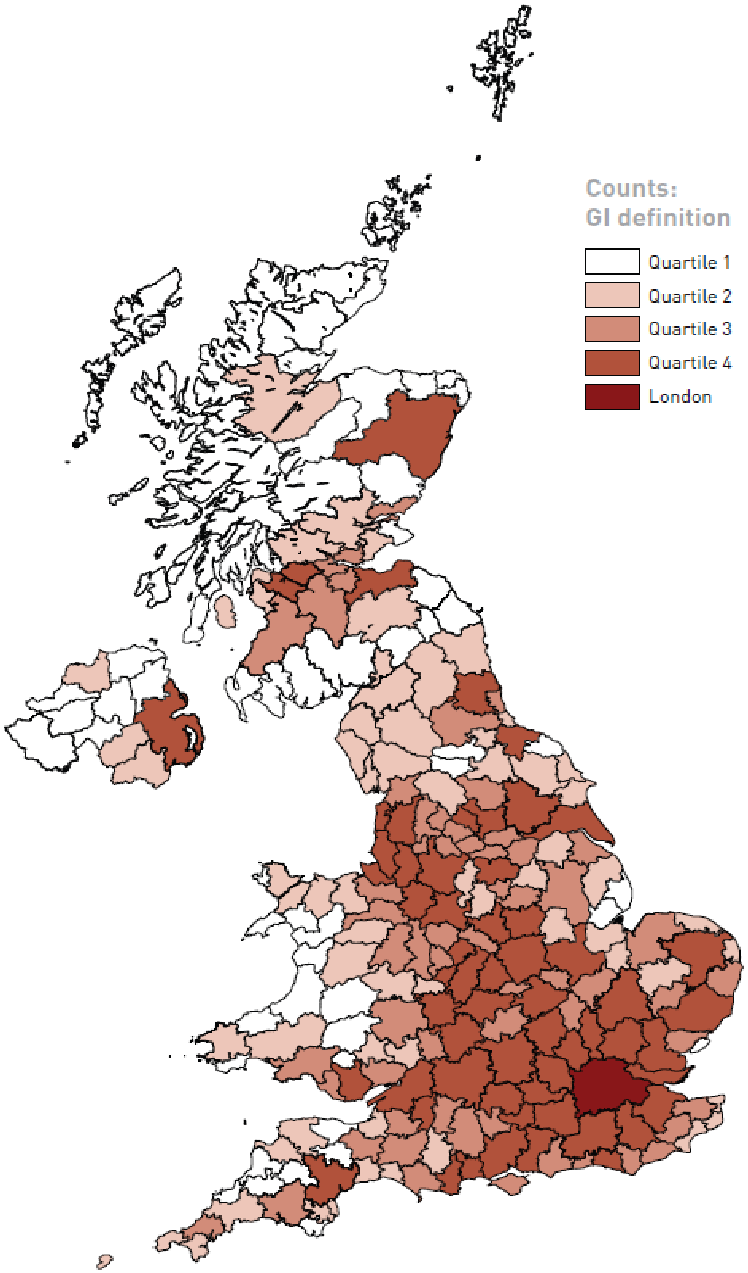
	2010-2012			
	Obs	**Mean**	**Median**	**% all jobs**
A. SIC07				
Other	70,805	21.13	4	96.23
ICT mf and services		17.06	2	3.77
B. GI sector/product				
Other	70,805	20.88	4	92.86
ICT mf and services		21.65	3	7.14

*Note: **sub-sample of firms reporting employment to Companies House.** Data is averaged over 2010-2012.*

... but employment data is a bit more mixed.

	2010-2012			
	Obs	**Mean**	**Median**	**% all jobs**
A. SIC07				
Other	70,805	21.13	4	96.23
ICT mf and services		17.06	2	3.77
B. GI sector/product				
Other	70,805	20.88	4	92.86
ICT mf and services		21.65	3	7.14

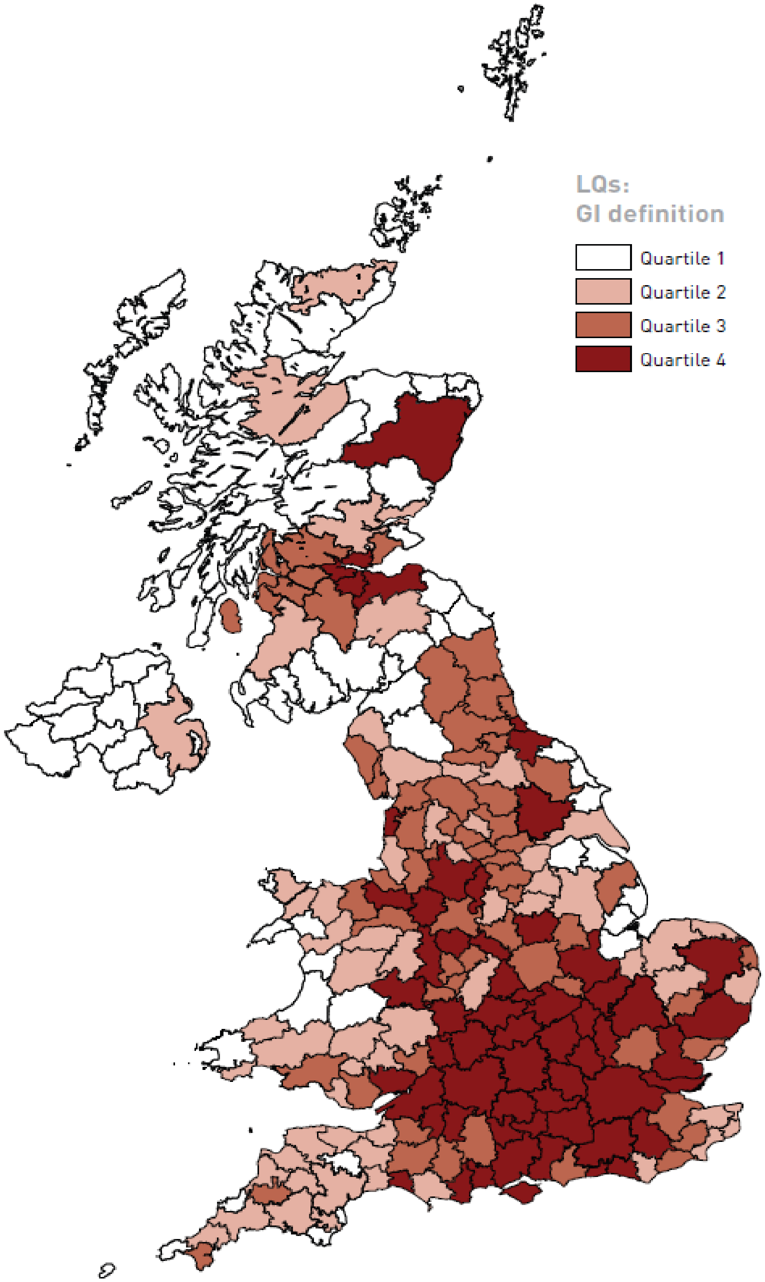
*Note: **sub-sample of firms reporting employment to Companies House.** Data is averaged over 2010-2012.*



Company counts are highest in **London**.

But we also find large counts in **Manchester, Birmingham, Bristol** and **Brighton** ...

... as well as the wider **Greater South East**.

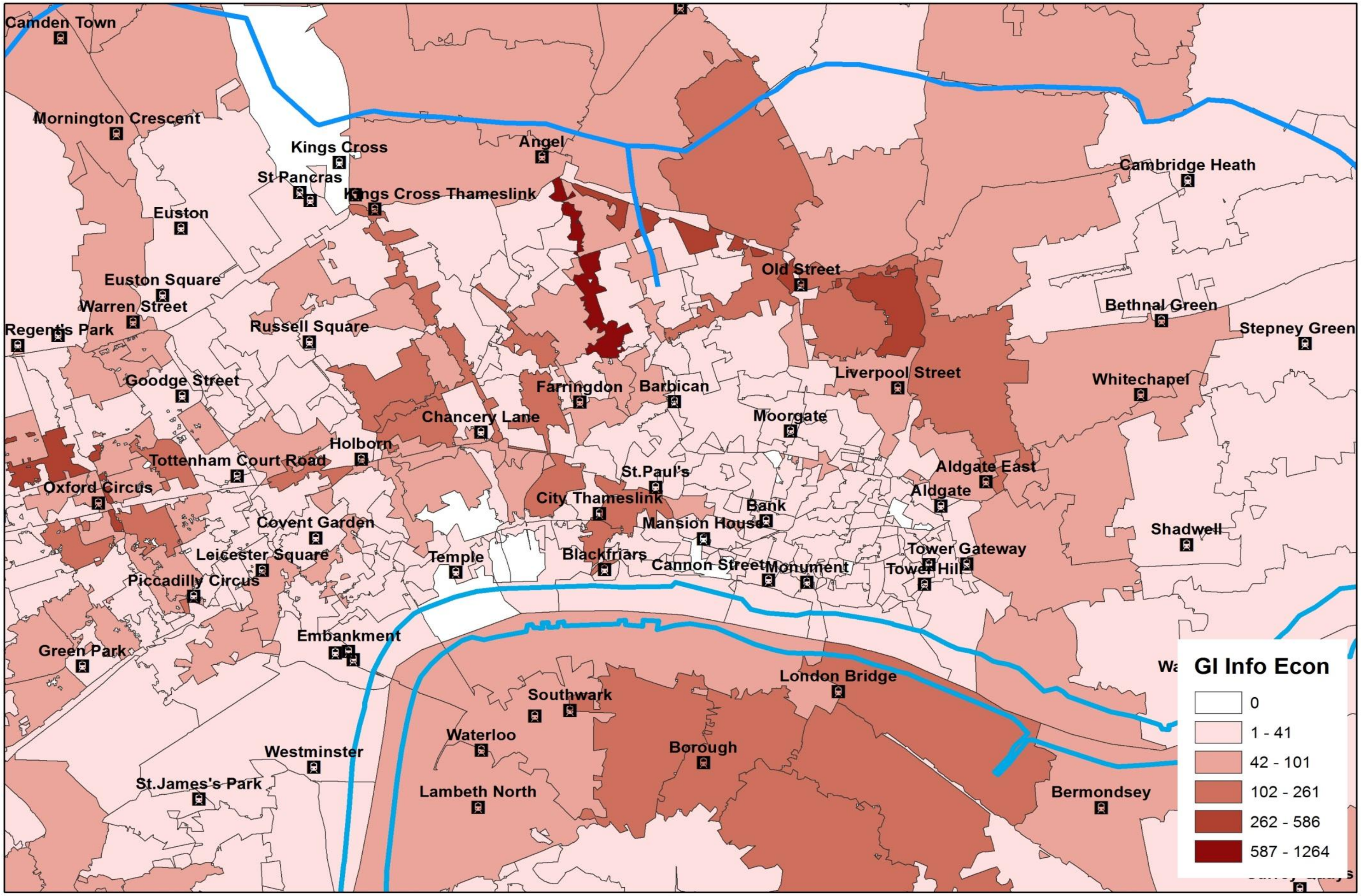


Location quotients measure local industry concentration.

The biggest ICT hotspots are in the **Greater South East** (tho **not London**) ...

... and we also find **Aberdeen**, **Blackpool, Coventry/Warwick** and **Middlesbrough** in the top 30.

IE firms in East London



Discussion

- **Alternative analysis of UK digital economy, using big data**
- Find (a lot) more ICT companies than SIC-based analysis
- Firms are larger, more established & resilient than popular perceptions; ICT diffusion into services and engineering
- Helpful input into national, local policymaking

- **Pros and cons of big data (at least in this context):**

- **Pros**: scale, speed of access, dimensionality and reach
- **Cons**: lack of structure, no metadata, fuzziness, validation

Thanks.

m.a.nathan@lse.ac.uk
@iammaxnathan

Robustness checks

	Observations	%
A. SIC 07 - manufacturing and services		
Other	**1,783,973**	**91.83**
Information Economy	**158,810**	**8.17**
B. Gi sector and product – manufacturing and services		
Other	**1,716,983**	**88.38**
Information Economy	**225,800**	**11.62**
C.SIC07 - services only		
Other	1,789,405	92.11
Information Economy	153,368	7.89
D. SIC07 - services, manufacturing & supply chain		
Other	1,748,607	90.01
Information Economy	194,176	9.99
E. Gi sector		
Other	1,637,606	84.29
Information Economy	305,177	15.71
F. Gi sector and product - manufacturing and services (0.5% threshold)		
Other	1,749,376	90.04
Information Economy	193,407	9.96
Total	1,942,783	100

Startups by postcode sector

		#Information economy startups		#information economy firms	
PC sector	#startups	SIC	sector-product	SIC	sector-product
EC1V 4	1557	194	242	1059	1277
BN36	368	200	218	784	860
N12 0	615	111	139	356	457
EC1V 2	288	92	106	533	590
HP11	190	95	96	488	518
SW19 1	254	77	86	363	413
CV12	293	75	85	389	440
W1B 3	370	70	83	289	345
BH12 1	486	77	80	328	340
EC2A 3	292	69	77	276	348
SO23 7	93	57	59	271	294
E14 5	147	47	53	216	238
DA14 4	162	42	52	145	189
EC1N 8	249	37	51	194	268
NG2 7	100	41	50	250	298
BN11	263	39	48	315	390
FY4 5	109	28	47	218	293
BH11	91	38	47	340	379
E14 9	216	46	45	244	267
W1G 9	557	32	44	216	300

Application II: Modelling company lifecycle events

What are ‘events’?

- News information from different web sources
- Sources: news aggregators (like ITBriefing, PRWeb) and some major news agencies like Reuters or Yahoo News.com: **2,643** different sources
- Information matched using company name, then classified into different event types

```
+-----+-------------------------+
| id  | event_type              |
+-----+-------------------------+
| 201 | alliance_joint_venture  |
| 206 | contract_awarded        |
| 207 | employee_hiring         |
| 209 | management_change       |
| 216 | merger_acquisition      |
| 220 | product_launch          |
| 221 | property_deal           |
+-----+-------------------------+
```

Example: ‘event’ information

company_id	525999
event_type_id	201
date	24/10/2013
fragment	conference.Ã¢ Another key collaborator working with Launch Tennessee on this yearÃ¢s Southland is AC Entertainment, co-creators and producers of the Bonnaroo Music and Arts FestivalÃ¢Â¢. The two organizations worked
source_name	ITbriefing
doc_title	Launch Tennessee and PandoDaily Announce Joint Venture to Produce 2014 Southland Conference
url	http://www.itbriefing.net/index.php?name=News&file=article&sid=486145

Source: Growth Intel

At the moment we have **300k+ raw events** covering **c.30k companies**. This will rise in the coming months, but coverage is fundamentally uneven.

Major data issues

- **Unstructured data: no sampling framework.** Not all companies have events and not all events are scraped

- **What does the 'event' really represent?**

- Selected sample of companies: event recording strongly correlated with company characteristics
- Quality of the information (accuracy, reliability of records reported)
- Farmed information copied from other websites

Simple cleaning steps

1. Cleaning the data: drop duplicates (event from the same source and day)

2. For each event we keep only one source

2a. Investigate the quality issues by using firms' characteristics: is this informative about the quality of the source?

Type of events

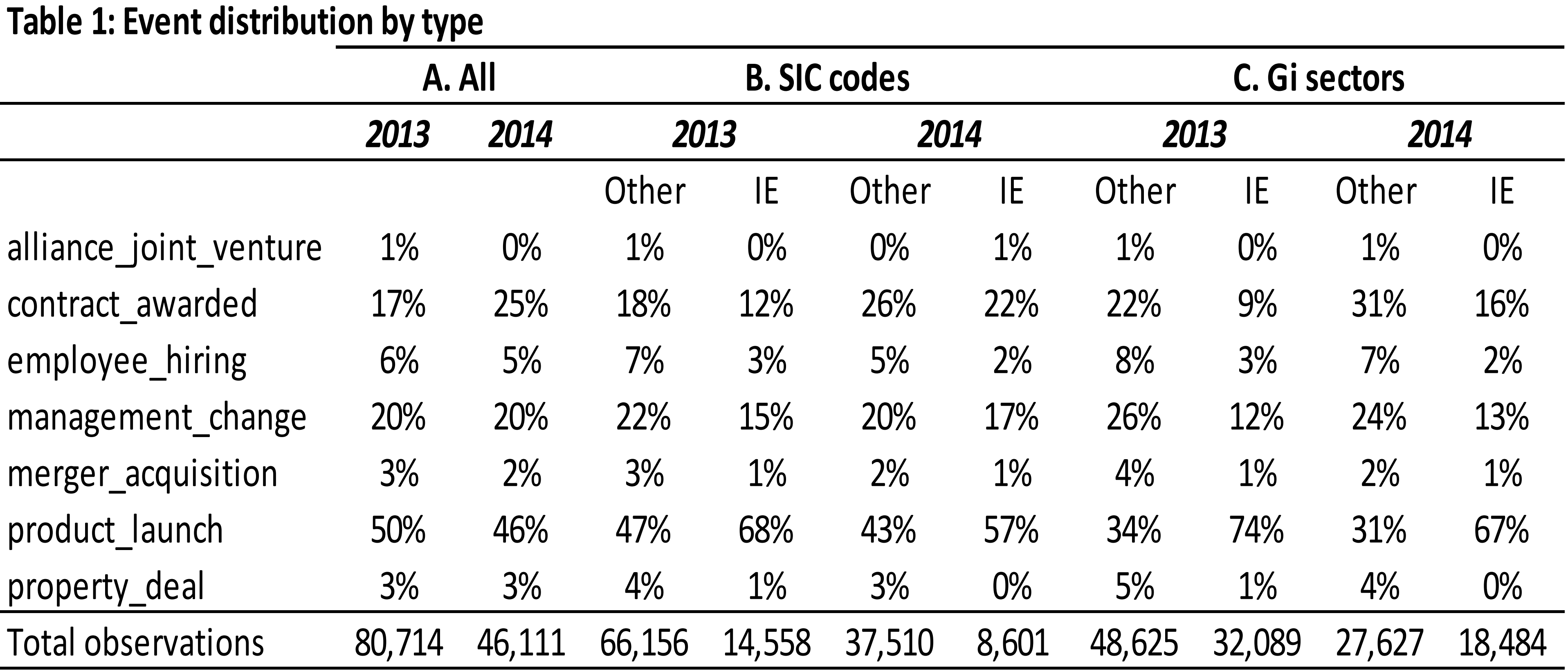
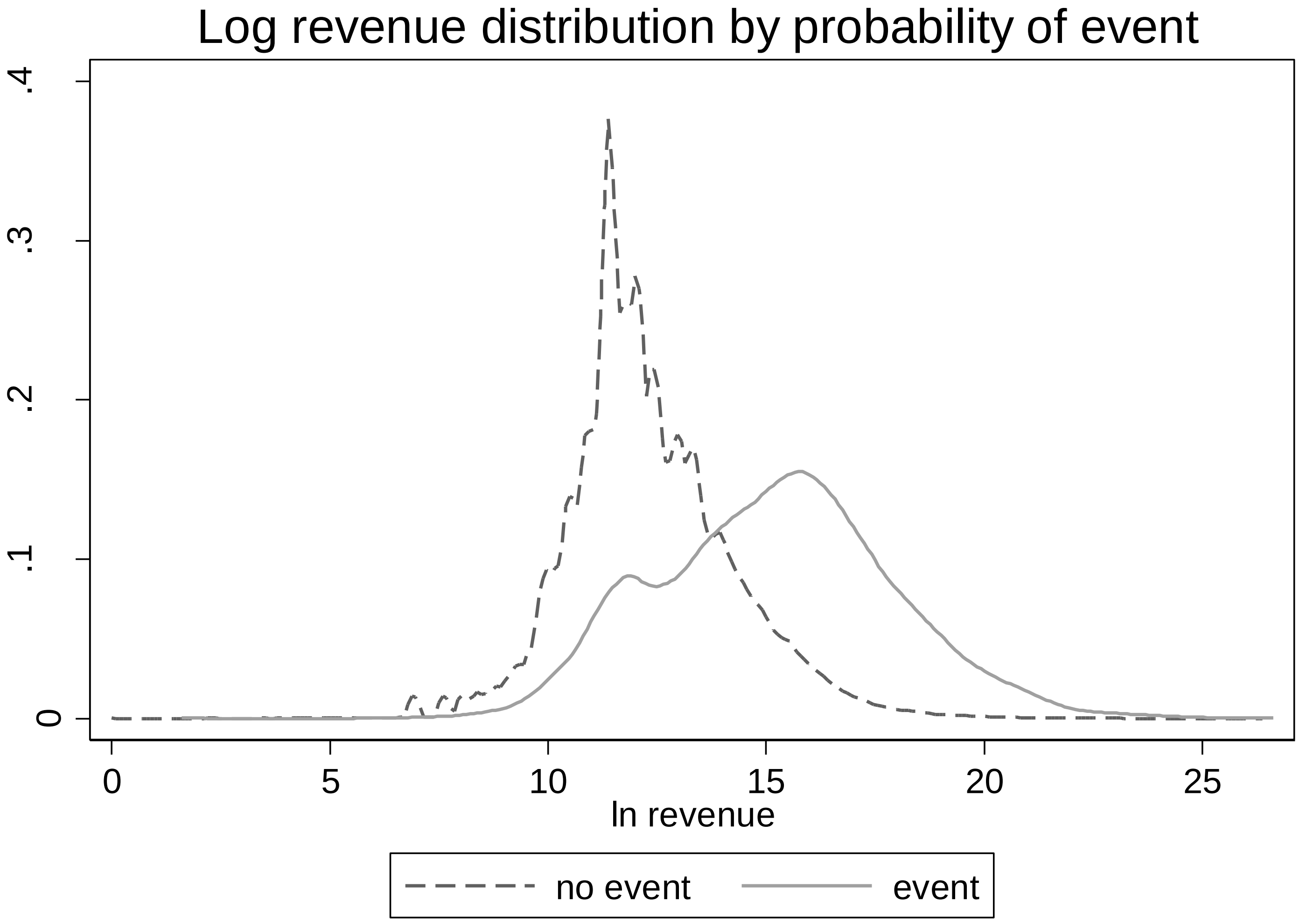


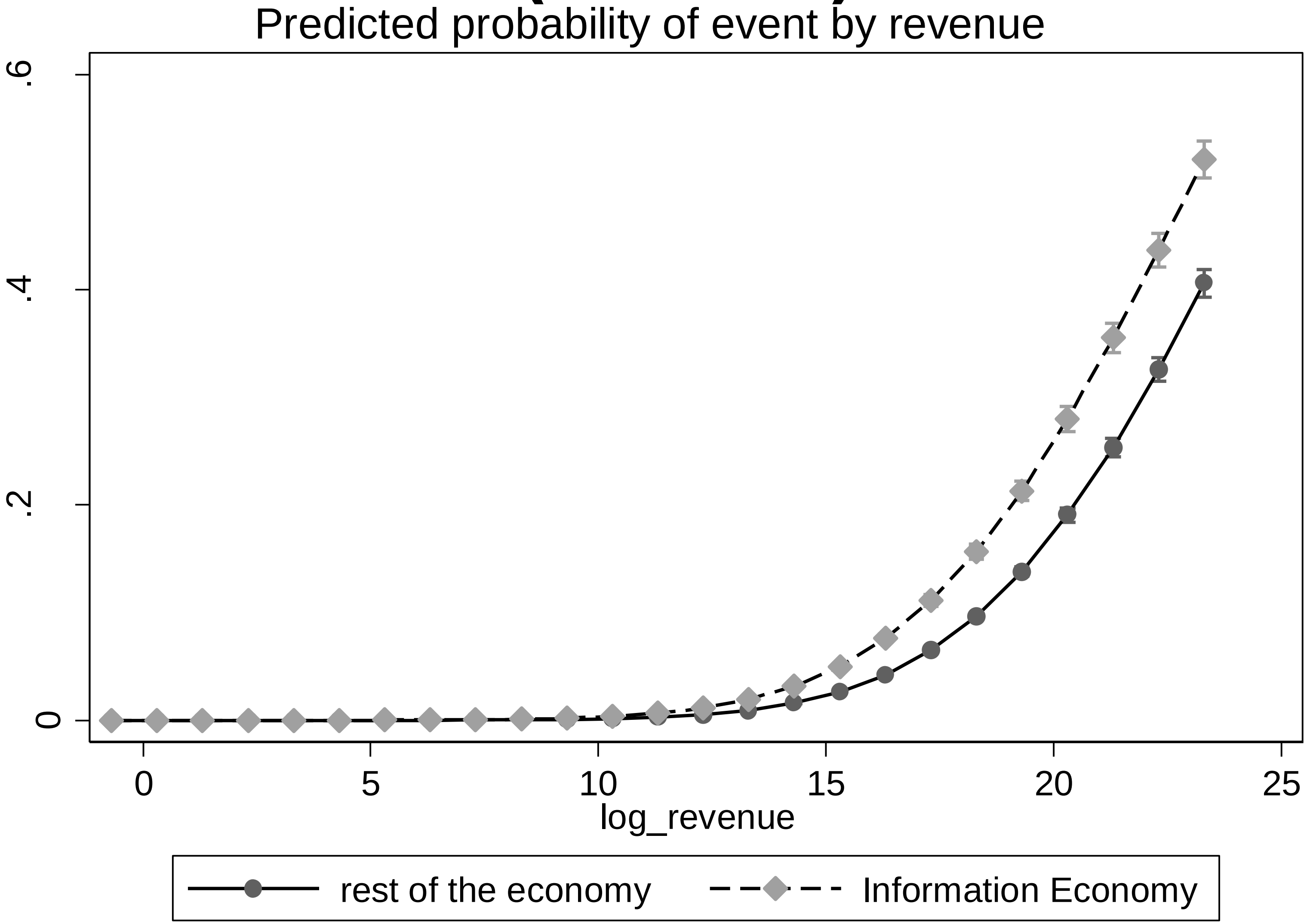
Table 1: Event distribution by type

	A. All		B. SIC codes				C. Gi sectors			
	2013	***2014***	***2013***		***2014***		***2013***		***2014***	
			Other	IE	Other	IE	Other	IE	Other	IE
alliance_joint_venture	1%	0%	1%	0%	0%	1%	1%	0%	1%	0%
contract_awarded	17%	25%	18%	12%	26%	22%	22%	9%	31%	16%
employee_hiring	6%	5%	7%	3%	5%	2%	8%	3%	7%	2%
management_change	20%	20%	22%	15%	20%	17%	26%	12%	24%	13%
merger_acquisition	3%	2%	3%	1%	2%	1%	4%	1%	2%	1%
product_launch	50%	46%	47%	68%	43%	57%	34%	74%	31%	67%
property_deal	3%	3%	4%	1%	3%	0%	5%	1%	4%	0%
Total observations	80,714	46,111	66,156	14,558	37,510	8,601	48,625	32,089	27,627	18,484

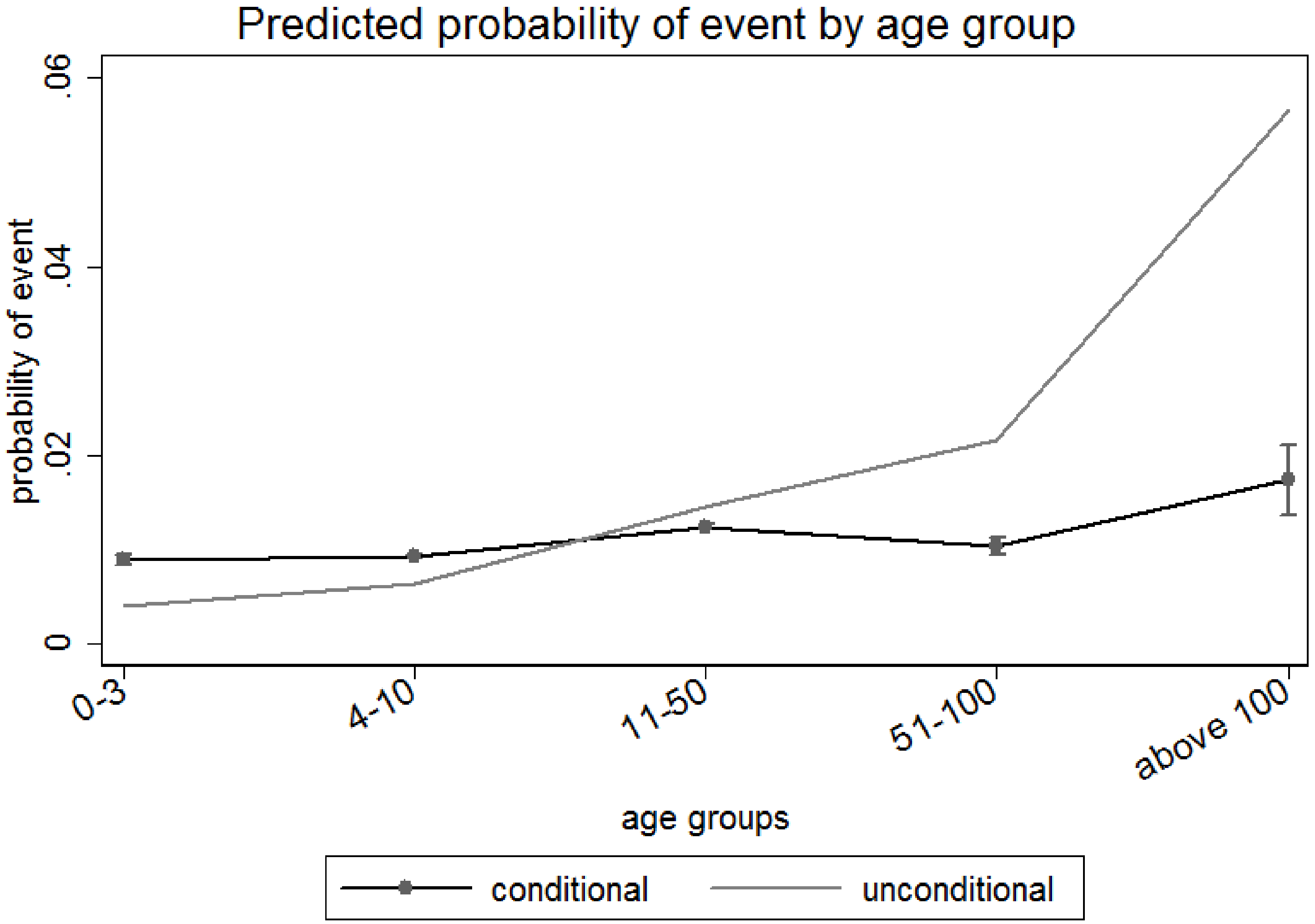
Company characteristics

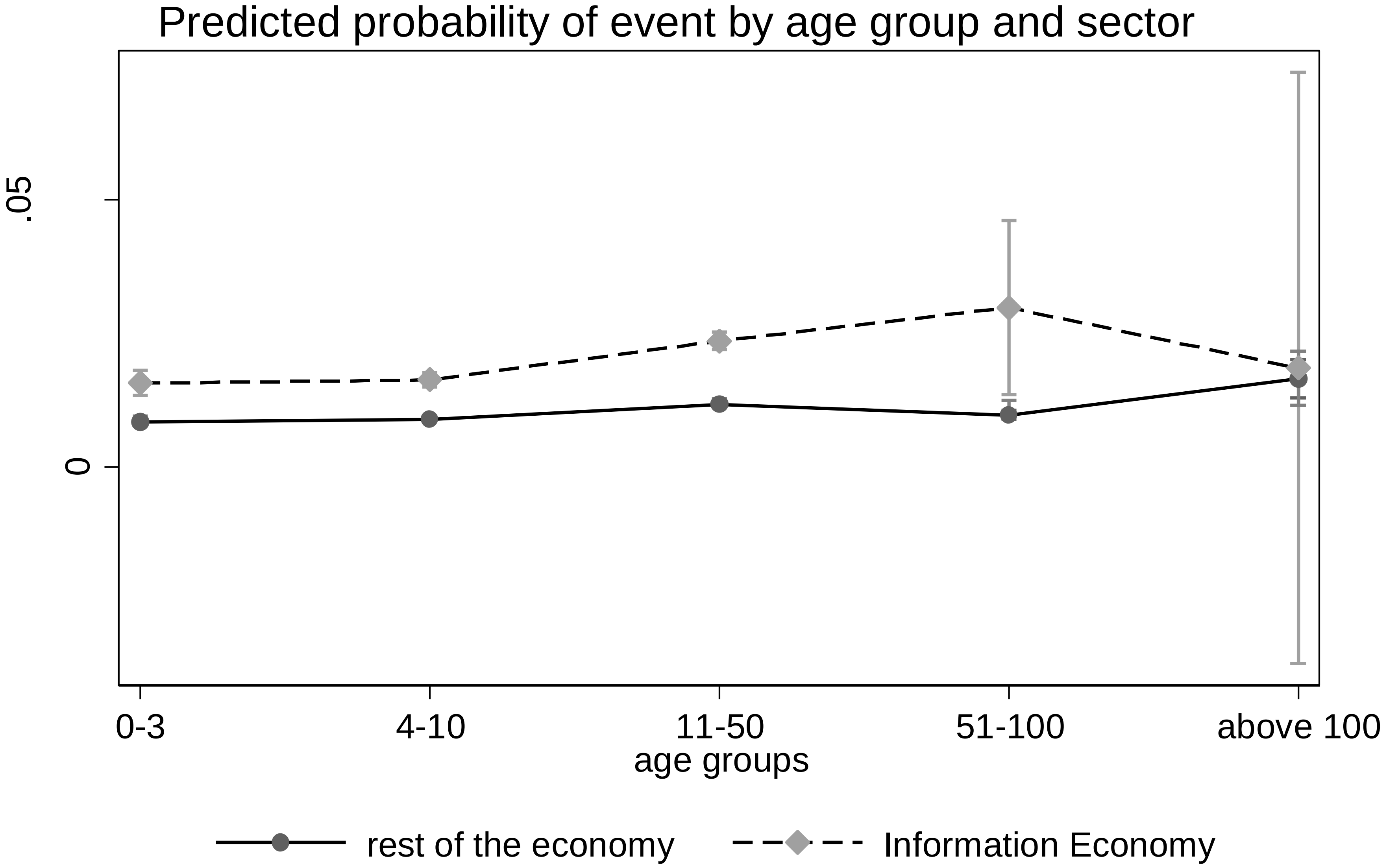


Company characteristics (cont'd)

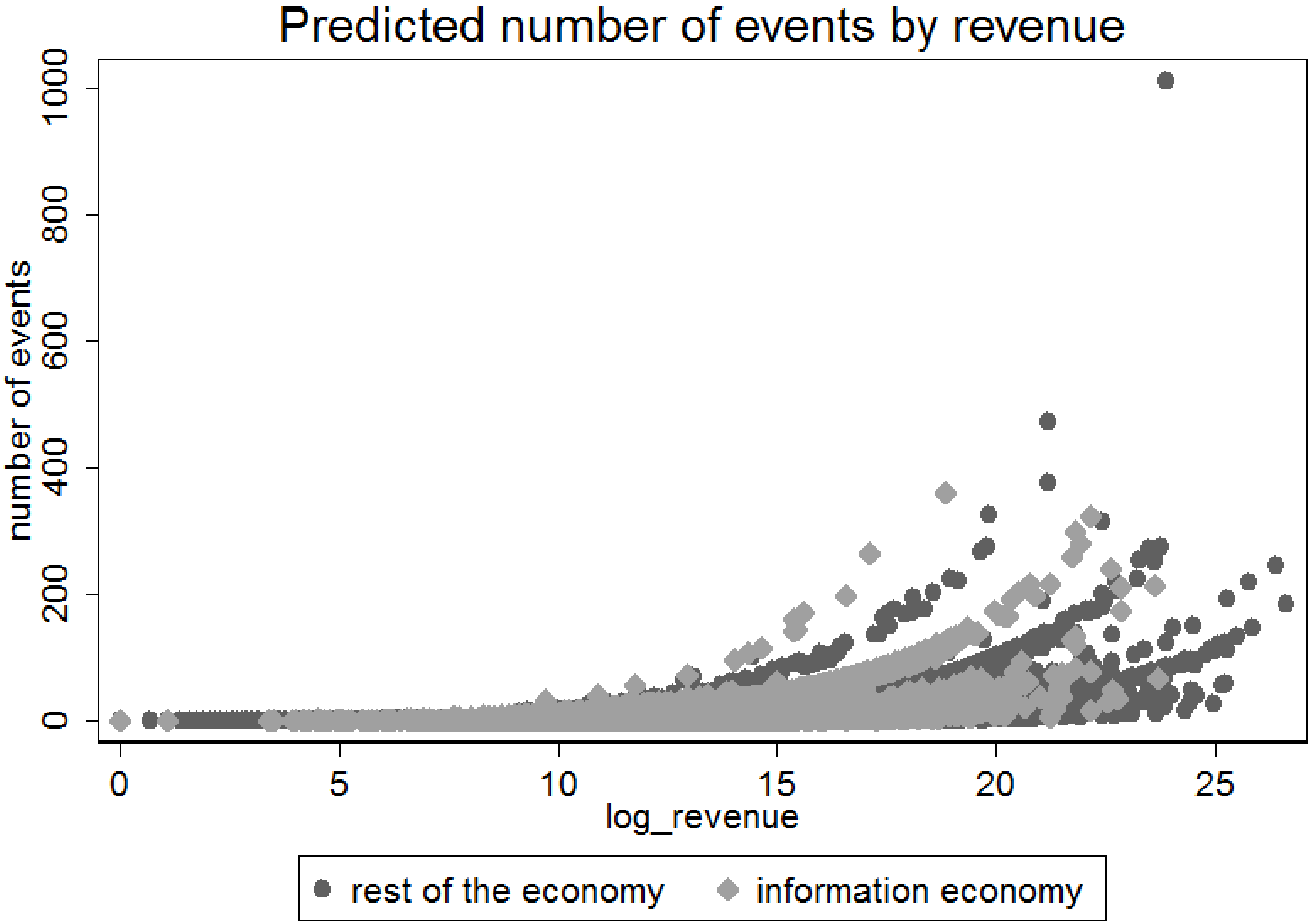


Company characteristics (cont'd)

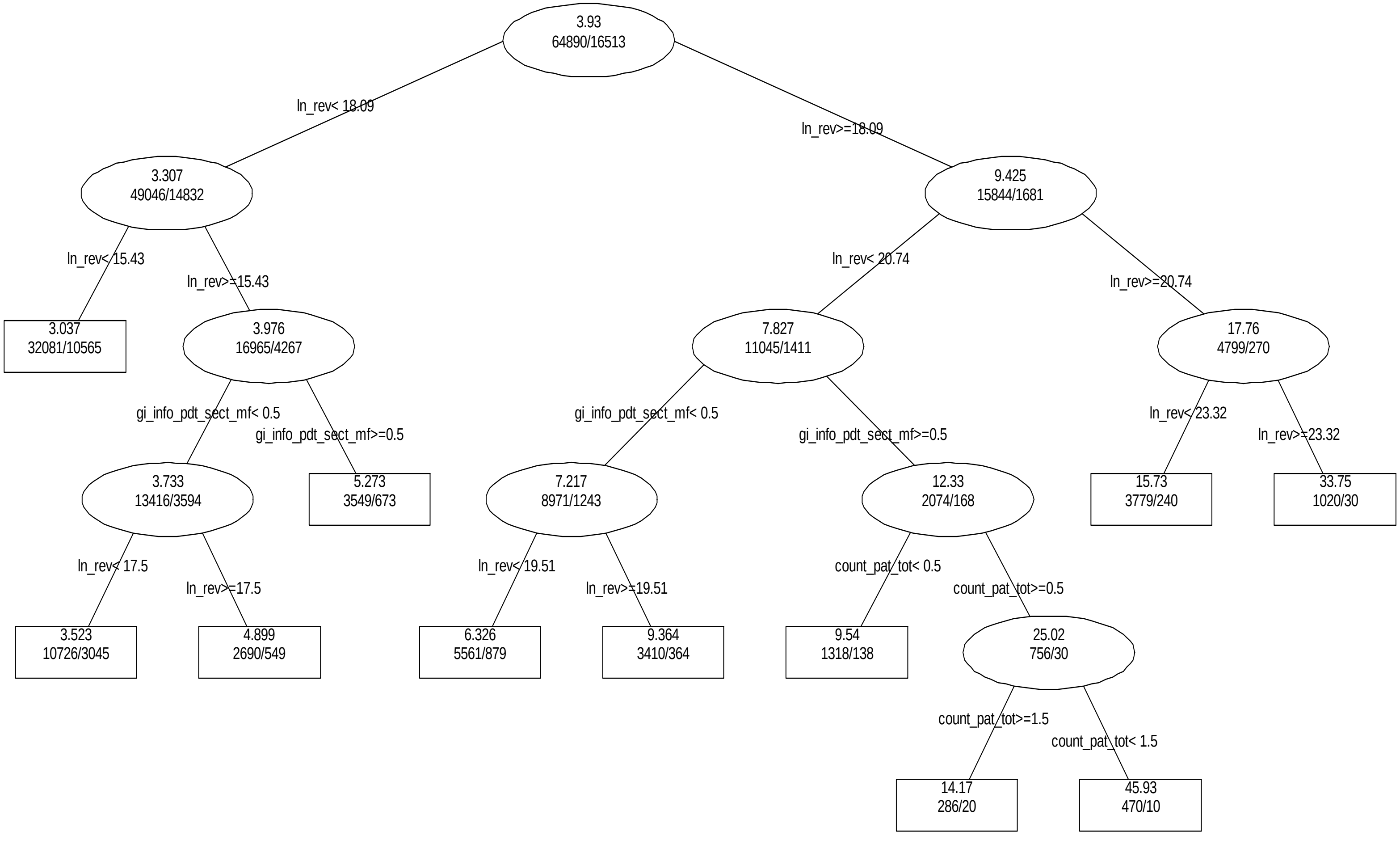



Predicted probability of event by age group and sector
.05
0
0-3
4-10
11-50
51-100
above 100
age groups
rest of the economy
Information Economy

Sector and event counts



Classification Tree for event counts


3.93
64890/16513
ln_rev< 18.09
ln_rev>=18.09
3.307
49046/14832
9.425
15844/1681
ln_rev< 15.43
ln_rev>=15.43
ln_rev< 20.74
ln_rev>=20.74
3.037
32081/10565
3.976
16965/4267
7.827
11045/1411
17.76
4799/270
gi_info_pdt_sect_mf< 0.5
gi_info_pdt_sect_mf>=0.5
gi_info_pdt_sect_mf< 0.5
gi_info_pdt_sect_mf>=0.5
ln_rev< 23.32
ln_rev>=23.32
3.733
13416/3594
5.273
3549/673
7.217
8971/1243
12.33
2074/168
15.73
3779/240
33.75
1020/30
ln_rev< 17.5
ln_rev>=17.5
ln_rev< 19.51
ln_rev>=19.51
count_pat_tot< 0.5
count_pat_tot>=0.5
3.523
10726/3045
4.899
2690/549
6.326
5561/879
9.364
3410/364
9.54
1318/138
25.02
756/30
count_pat_tot>=1.5
count_pat_tot< 1.5
14.17
286/20
45.93
470/10