**Technical Report for ESRC Project: Grant reference: ES/J008303/1**

**Sustaining Growth for Innovative New Enterprises: UK Firm Data**

**Prepared by M. Sensier (marianne.sensier@manchester.ac.uk)**

**Award Holders**: Professor Philip Shapira and Professor Alan Harding

**Researchers**: Dr Marianne Sensier, Dr Abdullah Gök and Dr Elvira Uyarra

**Database collection description:** To select the group of UK firms we initially searched in the FAME database (available from the University of Manchester Library) with keywords relating to the green goods sector, please see the publication Shapira, et al (2014, in Technological Forecasting & Social Change, vol. 85, pp. 93-104) for further details on the keywords. This database contains anonymized firm data from a sample of UK firms in the green goods production industry. We combine data from structured sources (the FAME database, patents and publications) with unstructured data mined from firm's web-sites by saving key words in text and summing up counts of these to create additional explanatory variables for firm growth. The data is in a panel from 2003-2012 with some observations missing for firms. We collect historical data from firm's web-sites available in an archive from the Wayback machine.

**Table 1: Description of UK Variables and Sources, panel 2003-2012**

**Data in file: UKFirmSI\_ESRC\_J008303.xls**

|  |  |  |
| --- | --- | --- |
| **Variable** | **Source** | **Description/Transformation** |
| firm | Own creation | Firm number between 1-304, anonymous (Firms removed no. 28, 161, 189, 211, 248, 263, 272, 292) |
| year | Own creation | Year for panel data, 2003-2012 |
| sector | Own creation | Sectors: 1=Environment/pollution control; 2=Building industry; 3=Battery; 4=Alternative vehicle/ Fuel; 5=Renewables; 6=Other. |
| grpats | Derwent | Total number of green patents received |
| grpubs | Web of Science | Total number of green published articles |
| start | FAME | Year of incorporation |
| SIC07 | FAME | Primary firm standard industrial classification code based on 2007 industrial categories |
| NUTS3 | Own creation | Location of firm in NUTS 3 European Geographical area |
| NUTS1code | Own creation | Location of firm in NUTS 1 European Geographical area: 1=North East; 2=North West; 3=Yorkshire & Humberside; 4=East Midlands: 5=West Midlands; 6=East; 7=London; 8=South East; 9=South West; 10=Wales; 11=Scotland; 12=Northern Ireland. |
| SizeDummy | Own creation | Size of firm by employment: 1=emp<10 (micro); 2=10<emp<50 (small); 3=50<emp<250 (medium); 4=emp>250 (large). |
| legalformcd | FAME | Ownership of firm: 1=Guarantee; 2= Private Limited; 3= Public AIM; 4=Public, not quoted; 5=Public, quoted. |
| ownership | FAME | Ownership of firm: 1=Independent; 2= Subsidiary of UK company; 3= Subsidiary of foreign company; 4=Holding company |
| CPI | ONS | UK consumer price index, 100 in year 2005 |
| turn | FAME | Firm Turnover (£ thousands) |
| realturn | FAME | turn/(CPI/100) |
| grlrealturn | FAME | 100\*(ln(realturn(t))-ln(realturn(t-1))) |
| grtposdummy | FAME | 1 when grlrealturn>0 |
| grt5dummy | FAME | 1 when grlrealturn>5% |
| grt10dummy | FAME | 1 when grlrealturn>10% |
| grt20dummy | FAME | 1 when grlrealturn>20% |
| mlogitdep | FAME | 1 when grlrealturn when<0; 2 when 0>grlrealturn<20%; 3 when grlrealturn>20% |
| emp | FAME | Total number of employees |
| lemp | FAME | log of employees |
| gremp | FAME | Growth in employees =(emp-emp(t-1))/emp |
| oturn | FAME | Overseas turnover reported in company accounts |
| oturnDummy | FAME | 1 when Overseas turnover is in company accounts |
| rnd | FAME | R&D spending reported in company accounts |
| RnDDummy | FAME | Dummy of 1 when R&D spending is reported in company accounts |
| HRST | Eurostat | People employed in Human Resources in Science and Technology as a % of the active population, by NUTS 2 region |
| HRSTUK | Eurostat | People employed in Human Resources in Science and Technology as a % of the active population for the UK |
| Skillsgap | Eurostat | HRST-HRSTUK |
| paidgrants | TSB | Total number of grants received by firm from TSB (grants) |
| Pubfundwebdummy | TSB | 1 when Paid Grants or pubfund web mentions are positive |
| **Web Variables** |  | **See Table 2 for Keywords** |
| prodstan | ICA | Counts of keywords on Product Standards |
| glinks | ICA | Counts of keywords on Government Links |
| pubfund | ICA | Counts of keywords on Public Funding |
| products | ICA | Counts of keywords on Products |
| manuint | ICA | Counts of keywords on Manufacturing Intensity |
| custmr | ICA | Counts of keywords on Customer Relations |
| custtn | ICA | Counts of keywords on Customisation |
| invnrw | ICA | Counts of keywords on Investment (narrow definition) |
| vencapnw | ICA | Counts of keywords on Venture Capital Investment |
| rndins7 | ICA | Counts of keywords on R&D Intensity |
| partner | ICA | Counts of keywords on Partnerships |
| ukcitiesins | ICA | Counts of keywords on UK City mentions |
| rowcitiesins | ICA | Counts of keywords on UK Rest of World City mentions |
| memorg | ICA | Counts of keywords on Membership of Organisations |
| unilinks | ICA | Counts of keywords on University Links |
| trls | ICA | Counts of keywords on Trials |
| green | ICA | Counts of keywords on Greenness |
| awards | ICA | Counts of keywords on Awards attained |
| demos | ICA | Counts of keywords on Demonstrations |
| renewins | ICA | Counts of keywords on Renewables sector |
| lowcarbonins | ICA | Counts of keywords on Low Carbon sector |
| envins | ICA | Counts of keywords on Environmental sector |
| pages | ICA | Counts of number of firm’s web pages |
| phrases | ICA | Counts of number of phrases on firm’s web-site |
| nprodstan | ICA | 100\*(prodstan/phrases) |
| nglinks | ICA | 100\*(glinks/phrases) |
| npubfund | ICA | 100\*(pubfund/phrases) |
| nproducts | ICA | 100\*(products/phrases) |
| nmanuint | ICA | 100\*(manuint/phrases) |
| ncustmr | ICA | 100\*(custmr/phrases) |
| ncusttn | ICA | 100\*(custtn/phrases) |
| invnrw | ICA | 100\*(invnrw/phrases) |
| nvencapnw | ICA | 100\*(vencapnw/phrases) |
| nrndins7 | ICA | 100\*(rndins7/phrases) |
| npartner | ICA | 100\*(partner/phrases) |
| nukcitiesins | ICA | 100\*(ukcitiesins/phrases) |
| nrowcitiesins | ICA | 100\*(rowcitiesins/phrases) |
| nmemorg | ICA | 100\*(memorg/phrases) |
| nunilinks | ICA | 100\*(unilinks/phrases) |
| ntrls | ICA | 100\*(trls/phrases) |
| ngreen | ICA | 100\*(green/phrases) |
| nawards | ICA | 100\*(awards/phrases) |
| ndemos | ICA | 100\*(demos/phrases) |
| nrenewins | ICA | 100\*(renewins/phrases) |
| nlowcarbonins | ICA | 100\*(lowcarbonins/phrases) |
| nenvins | ICA | 100\*(envins/phrases) |

Sources: ICA is IBM Content Analytics this collates the text mined from firm’s web-sites.

ONS: Office for National Statistics (<http://www.ons.gov.uk/ons/rel/cpi/consumer-price-indices/index.html>)

FAME: Bureau van Dijk, (2004). Advanced Internet Quick Guide. Bureau van Dijk Electronic Publishing.

Eurostat: <http://ec.europa.eu/eurostat/data/database>

Note: “ln” in the grlrealturn transformation is taking the natural logarithm of a number; “t” is a time operator for the current period “t-1” is for the previous time period.

**Table 2: Keywords used in the webscraping process**

|  |  |
| --- | --- |
| **Variable** | **Keywords Used** |
| products | buy, order, sale, sell, brand, model, version, catalog, product |
| trls | experimental, experimentation, exploratory, pilot, prelim, preliminary, provisional, tentative, test, testing, trial, in-process |
| demos | demo, affirmation, confirmation, exhibition, exposition, expression, illustration, presentation, proof, showing, substantiation, test, validation |
| rndins7 | lab, laboratory, research, development, R&D,researcher, scientist |
| manuint | manufactur produc assembly, \*fabricat\*, machined, machinery, machines, mold, prefab, process, construct, weld, engineer |
| vencapnw | venture capital, VC, valuation, public offering, IPO, acquisition, merger, liquidity event, seed funding, round of funding, stage of funding |
| invnrw | investor, investment firm, private equity, bank, adventure capitalist, bridge loan, buyout, corporate venture capital, corporate venturing, deal flow, debt financing, direct financing, drive-by deal, due diligence, equity financing, financier, full ratchet, fund of funds, institutional investors, lead investor, leveraged buy-out, liquidity event, lock-up period, management buy-in, management buy-out, master limited partnership, mezzanine debt, mezzanine financing, owner-employee, portfolio company, private equity, private placement, raising capital, recapitalization, resyndication, risk capital, investment company, secondary purchase, syndication, term sheet, underwriter, vulture capitalist |
| glinks | ((^| )(DEFRA)($| ))|(department( )for( )environment), governement[a-z]\*, public[a-z]\*( )(sector[a-z]\*|organisation), council, european( )commission, (^| )TSB($| ),technology( )strategy( )board, parliament |
| unilinks | university, college, institute, academ |
| ukcitiesins | Look for names of UK cities and town in firm-level website data. |
| rowcitiesins | Look for names of non-UK cities and town in firm-level website data. |
| partner | partner, stakeholder, distributor, collaborat, alliance, joint venture, agreement, supplier, parts manufacturer |
| green | Green, sustainabl\*, eco-friendly, environmentally friendly, environmentally conscious, renew\*, clean air, clean coal, clean energy, clean fuel, clean technology, cleantech, climate change, global warming, rising temperatures |
| custtn | Key terms: custom, made to order, made-to-order, bespoke, tailor |
| awards | award, prize, certif\*, |
| memorg | member, affiliation, association, club, participation, network |
| prodstan | (^| )ISO[0-9]+, (^| )ISO( )[0-9]+, (^| )BS[0-9]+, (^| )BS( )[0-9]+, comply[a-z]\*, complie[a-z]\*, compliant[a-z]\*, accordan[a-z]\*, green( )deal, (product[a-z]\*|green|UK|EU|Union|British)( )standard[a-z]\*, regulation[a-z]\*, directive[a-z]\* |
| custmr | Customer, consumer, end user, purchaser, client, customer-base, buyer |
| pubfund | (government[a-z]\*|public|state|EU|Union|RDA|DEFRA|programme[a-z]\*)( )(funding[a-z]\*|subsidy[a-z]\*|grant|grants|loan|support) |

Note: 2012 data was from Firm’s web-site, historical data was scraped from firm’s archived web-sites on the Wayback machine.

**Database collection method**: We collected the financial information on the UK firms by downloading Companies House data from the FAME database available through the University of Manchester Library (see <http://www.library.manchester.ac.uk/searchresources/databases/f/>). Grant information on companies came from the Technology Strategy Board. Patent information was from the Derwent database and publication information was from the Web of Science. The Consumer Price index was from the Office for National Statistics (<http://www.ons.gov.uk/ons/rel/cpi/consumer-price-indices/index.html>). The Human Resources in Science and Technology variable was from the Eurostat database (<http://ec.europa.eu/eurostat/data/database>).

Unstructured data was mined from firm's web-sites. The UK Intellectual Property Office has clarified that the data mining we are doing and the way we are doing it is permissible. See: <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf>

**Data sourcing, processing and preparation**: Data on patents, publications and grants were linked to the relevant firms. Unstructured data was mined from the firm's live web-sites and from archives on the Wayback machine. The data was collated in IBM Content Analytics software in XML format. The data was then imported into Vantage Point software where text fields were merged and the data cleaned. Variables were extracted from Vantage Point and imported into Stata 12 for analysis. Data transformations were undertaken in Stata 12 and the description of these are listed in column 3 of Table 1 in DataTechnicalReportESRC\_SI (above in this document). We standardise the web variables by dividing them by the number of phrases counted across all the firm’s web-pages for the year and then multiply this by 100.