Rainer Frietsch

Fraunhofer ISI, Karlsruhe, Germany

Combining Databases for Forecasting Purposes

– Patents and further economic data –

Presentation at the

WIPO-OECD Workshop on the Use of Patent Statistics

Geneva, 11th/12th October 2004

Fraunhofer Institute Systems and Innovation Research

Structure of presentation

1. Introduction
2. Random sample of applicants
   - Drawing the sample
   - Adding further information
   - Descriptive Statistics and Outlook
3. Large companies – large applicants
   - Data sources
   - Descriptive Statistics and Outlook

1. Introduction

Why forecasting?

Governments or public authorities:	"early warning system", monitoring, "governance"
Scientific research:	future structures, consulting public or private authorities, scientfic curiosity
European Patent Office:	Planning future workload and future budgets

1. Introduction

There are two ways of assessing the problem:

1. Macro approaches: The units of the analysis are countries, sectors (or product groups)
   - Advantage: Publicly available data can be used
   - Problem: Concordance between technologies (patents) and sectors is needed

2. Micro approaches: The unit of analysis are companies/applicants (or applications)
   - Advantage: Individual idiosyncrasies can be taken into account
   - Problem: Availability of (appropriate) data

One can think of a third way: Setting up a simulation, which uses and produces artificial data

---

2. Random sample of applicants

Why (another) micro approach?

Usual micro approaches: a) Existing surveys containing patent information, like the CIS or ad-hoc/occasionally conducted surveys

b) Applicant Panel Survey of the EPO (similar efforts are undertaken by the USPTO or the JPO, for example)

Shortcomings/disadvantages of these surveys:

Ad a) - Analysing patent filings is not in the centre of these studies
- Though only innovative companies are surveyed, only about 30% use patents at all
- EPO patents (and also not PCT vs. Euro-direct) cannot be distinguished

Ad b) - Expensive task
- unit non-response of 60-70% (+ item non-response)
- R&D-centred, restricted additional information

2. Random sample of applicants

There is still a need for further micro data

- Aim: database that is cheap, easy applicable and easy replicable

  -> Combining databases

Method:

- Draw a random sample of applicants at the EPO (of the year 2000)
- Pool the data with further information from (publicly) available firm databases
- Analyse the dependency of filings on other (economic) factors
- *(Use this knowledge for forecasting future filings)*

2. Random sample of applicants

Drawing the sample:

- Highly skewed distribution of applicants -> pure random sample is not appropriate
- Applicant panel survey of the EPO: sampling of applications
- USPTO and JPO surveys: ???
- We decided to use a disproportionally stratified sampling approach
  - to cope with the skewness
  - to reduce the heterogeneity (and sampling error) in the sample

2. Random sample of applicants

- As strata, we used the size of applicants in terms of applications
- We drew a sample of 1,000 applicants with the following groups:
  - 300 applicants with 1 patent
  - 300 applicants with 2-3 patents
  - 300 applicants with 4-30 patents and
  - 100 applicants with more than 30 patents
- BUT: sample is not representative for the "real" distribution of applicants
  - ➢ re-distribution to make statements about the population of the year 2000
- The weights for this procedure are calculated as: ("real" N) / (sampling N); for each size group and each country of origin
  - ➢ -> 2 in 1 (re-distribution + projection to population)

2. Random sample of applicants

- In the next step, we added further information about the applicants from publicly available databases namely:
- D&Bs "World Base" and "Market Europe", Creditreform (AT, CH, DE) and Hoppenstedt (DE), from which we received the following information:
    - sector: 923 companies (93.0%)
    - employees: 672 companies (67.7%)
    - sales: 558 companies (56.3%)

2. Random sample of applicants

Applicants and applications by number of employees

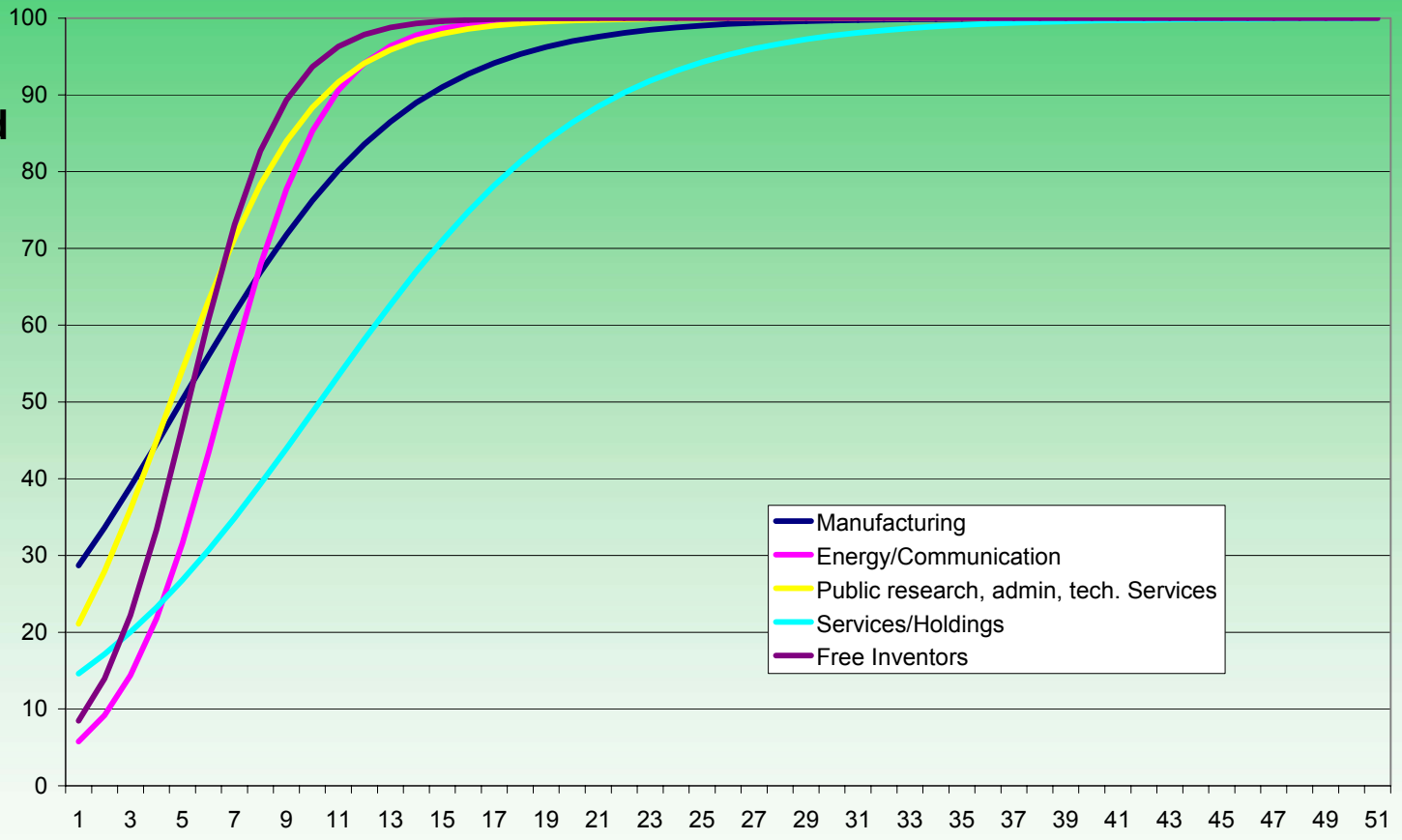
Source: EPO Sample; Fraunhofer ISI calculations.

	Applicants			Applications		
	N	%	valid	N	%	valid
1-9	4769	10.4	19.2	7843	5.5	7.0
10-49	4879	10.6	19.6	8776	6.1	7.9
50-249	6445	14.0	25.9	11871	8.3	10.6
250-499	2196	4.8	8.8	6037	4.2	5.4
500-1999	2958	6.4	11.9	14575	10.1	13.1
2000+	3603	7.8	14.5	62418	43.4	56.0
Total valid	**24851**	**53.9**	**100.0**	**111520**	**77.6**	**100.0**
not classified	21214	46.1		32210	22.4	
Total	**46065**	**100.0**		**143730**	**100.0**	

2. Random sample of applicants

Probabilities of having had an application in 1999 by sector and number of filings in 2000

Source: EPO Sample; Fraunhofer ISI calculations.



Fraunhofer Institute Systems and Innovation Research

2. Random sample of applicants

How to use all this for forecasting?

- Step1: Identify continuous and discontinuous applicants
- Step2: Estimate the N of filings in the following year for the continuous applicants with the help of a (e.g. Poisson) regression
- Step3: Draw a sample of the following year and apply the estimated coefficients to it

Filings in the following year = Filings in the base year X growth rate + N of filings of discontinuous applicants

Assumptions: Connection between independent variables and the dependent variable (patent filings) is more or less stable over time and the number of discontinuous applicants is stable, too

2. Random sample of applicants

Outlook

- Database can be replicated for any actual year (no historical data on firms is available)
- Approach can be applied to any subsample of applicants (e.g. the 14 technical clusters)
- Improvements might be possible by using more factors for stratifiying the sample (countries, technologies)
- Simple linear regression approach showed good results in estimating filings with the help of country, N of employees, sector and filings in the (previous) preceding year.
- More elaborated methods (like Poisson or negative binomial regression) may improve results.
- Knowledge about probability of continuous vs. discontinuous applicants together with the estimation of the amount of filings can be used for forecasting.

3. Large companies – large applicants

- Thus, the random sample approach leads to good data and promising results, another approach might bring further insights
- Especially the fact that no information on R&D is available is a disadvantage of this data
- Besides this, the fact of the skewed distribution of applicants in terms of applications suggests an "exclusive" look at the most relevant applicants

This is what we tried to do with the following approach

3. Large companies – large applicants

- The DTI (Department of Trade and Industry) of the British Government annually publishes a list with the 500 (600, 700) largest companies in terms of R&D expenditure (recently also including US grants in nine sectors; and in the future maybe even patent citations)
- By processing data between 1991 and 2000, we have been able to set up a database with 652 companies
- Mergers within the period are treated as one company over the whole period
- This lead to a data set, which contains R&D investment, number of employees, sales and profits, country of origin, sector (and some more variables, e.g. market capitalisation).
- In the next step, we were looking for their patent applications at the EPO in the 1990s

Problems: - not all companies use the groups name for applications (subsidiaries / affiliations)
- not all companies could be found in the database
- in some cases not all relevant patents could be identified

---

3. Large companies – large applicants

Correlations between N of patents and R&D expenditure, 1995 and 2000

Year of patent application	R&D-expenditures					
	2000	1999	1998	1997	1996	1995
1995	0.484(**)	0.476(**)	0.478(**)	0.474(**)	0.448(**)	0.411(**)
1996	0.538(**)	0.531(**)	0.523(**)	0.508(**)	0.460(**)	0.437(**)
1997	0.537(**)	0.529(**)	0.515(**)	0.490(**)	0.439(**)	0.412(**)
1998	0.540(**)	0.533(**)	0.510(**)	0.476(**)	0.427(**)	0.399(**)
1999	0.538(**)	0.530(**)	0.503(**)	0.470(**)	0.412(**)	0.392(**)
2000	0.528(**)	0.507(**)	0.484(**)	0.462(**)	0.408(**)	0.382(**)

** Correlation is significant at the 1% level

Source: DTI-Scoreboards, EPO; Fraunhofer ISI calculations.

3. Large companies – large applicants

Results

- Slightly more than 1% (only those, which we could identify in the EPO database) of the applicants in our database account for about 38% of all applications
- if one of these companies changes its efforts, activities, portfolio or just its habits, the total development of patent applications – especially in some technological fields – may be influenced massively
- R&D is an important, but not exclusive driving force in the application process
- We found some evidence for the thesis that patent applications lead to subsequent R&D investment
- It is appropriate not to restrict but to focus ones efforts on the largest applicants at the EPO not only because they account for a very large share of all applications but also because in many respects they act as trend setters also for the smaller companies.

---

3. Large companies – large applicants

Outlook

- The DTI increased the number of companies covered by its list -> higher N in the sample
- A “real” panel might be set up and individual developments over time can be taken into account
- An improvement of the identification of these companies in the EPO database may improve the results substantially
- Non-linear regression and methods of panel data analysis promise further insights
- The usefulness of this approach for “cluster analysis” is worth to be discussed

Thank you for your patience

Fraunhofer Institute
Systems and
Innovation Research