Academic patents in France: evidence 2006-2011

Lorenzo CASSI et Ibrahima WANE

APE-INV Final Conference
Paris 3rd September 2013

CONTEXT

- IPERU
  - Yearly project
  - financed by MESR
  - aim: to provide *consistent* indicators of university production in terms of publication and patents
  - main user: actor (e.g.. University itself) and policy maker
  - public available

- Last year: we provided a first estimation of academic patents for a set of 7 universities participating to a pilot-project

- This year: estimation based on previous (updated!) model of academic patents for all French universities

LIST OF CONTENTS

- A two steps presentation:
  - Retrieval methodology:
    - Data:
      - OST-PatStat (April 2012) / University Staff list (MESR)
      - TTO validation (7 universities)
    - Statistical model: estimation and prevision
  - First exploration in terms of applicants
    - Public vs. Private
    - National (local) vs. foreign

THREE STEPS METHODOLOGY

- Three steps procedure (Raffo and Lhuillery, 2009):

- *Parsing*: the standardisation and cleaning of two lists:
  - Academic (permanent) staff of French universities (MSER)
  - French inventors in OST-PatStat (April 2012), three offices: FR, EP and US (homonyms issue treated, Carayol and Cassi, 2009)

- *Matching*: semantic matching between the two lists (token similarities)

- *Filtering*: criteria that allow determining if observed matches identify the same person or not

FIRST TWO STEPS

1. **Data, cleaning and standardistaion of the two lists:**

- **Academic**. The list includes:
  - teachhing and reseracher staff of French Univerisities (UMR as well!) during the years 2004-2009
  - only permanent poistions personal
  - data were provided by the MESR and include, among other information, the first and last name, status, disciplinary section, date of birth
- The list of **patents** comes from the OST data base constructed from the base Patstat:
  - all inventors who have a home address in France and who had participated in a patent filed at the EPO, the USPTO and the INPI in 2004-2009;
  - Data processed to solve the problem of "who's who" in order to provide a reliable identifier inventors.

2. **Matching.** The two lists were matched on the basis of the names of inventors and researchers (semantic similarity). Obtained a set of XXX pairs, among them we have a subset of XXXX manually checked observation (i.e. staff of the 7 univeristies)

FILTERING

3. Goal: to **identify** automatically **academic inventors**

- We use a statistical model to estimate the probability that the matching between academic staff and inventor is correct. Two steps:

  1. Given the validation of matching made by 7 universities partecipating in the project, we estimate a set of explanatory variables available in both lists (e.g. MESR and PatStat).

  2. We use the estimates obtained to predict the probability that a matching is correct on all of the reference population.

TTO VALIDATION

- A member of the TTO of university of the sample has examined each matched inventor/academic staff in order to understand whether the observed matching was a correct or not:
  - 1 if the matching is correct, or
  - 0 if the *matching* is wrong

- Doing so, we get XXXX couples of inventor/academic staff that can be used for the estimation of the statistical model

STATISTICAL MODEL : ESTIMATION

- We specified a logistic regression to estimate such predictive ability of the following variables:
  - name (score calculated based on its frequency in Patstat);
  - similarity between the two names;
  - applicant (non-zero score if the university is the applicant, calculated based on its frequency in Patstat score);
  - age of the inventor at the time of publication of the patent;
  - correspondence between the scientific section of the researcher and the classification of technological fields (probability estimated from previous analysis, Patrick Llerena and couathors).

- Given the relatively small sample size, we used a bootstrap technique in order to not depend on the characteristics of the sample.

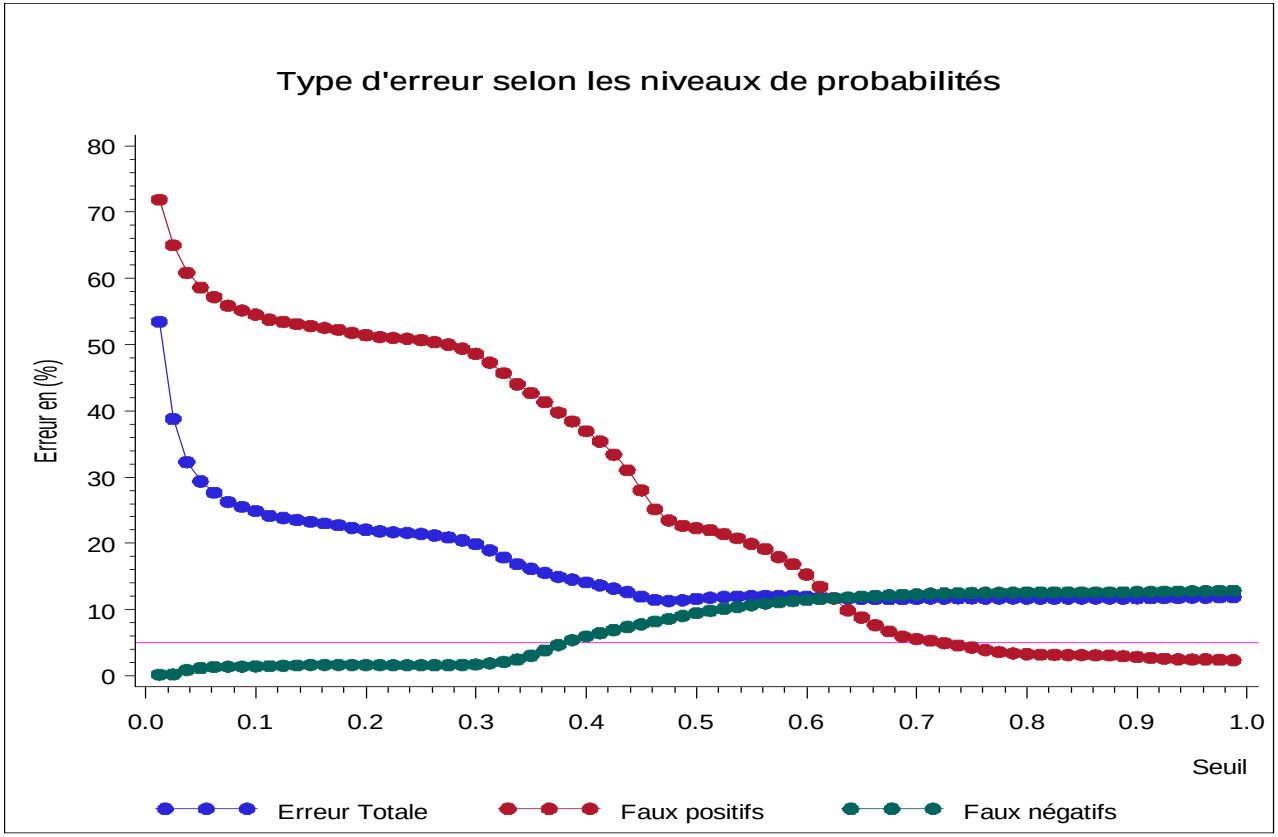
STATISTICAL MODEL : ERRORS

Our estimation based on our sample of XXXX couples allows us to quantify the prediction errors of our model.

	MODEL	inventor is not recognised as academic	inventor is recognised as academic
VALIDATION			
inventor is not an academic		No mistakes / True Negative (TN)	**False Positive (FP)**
inventor is an academic		**Faux Negative (FN)**	No mistakes / True Positive (TP)

STATISTICAL MODEL : ERRORS

The types of error (as a percentage) as function of the probability for which the matching between inventor and acdemic is estimated correctly



CHOOSING THRESHOLD

- To determine the acceptable probability (and hence the error rate), several criteria are possible.
- Following discussion with the group of the pilot universities, it was decided to identify a probability that allows "minimize" false positives.
- In this way, there is no risk of overestimating the role of higher education institutions in innovation: the estimate therefore provides a lower limit of the phenomenon.
- A false positive rate of 5% seems to be a good compromise errors of this type are at an acceptable level without causing a false negative rate and total error too high (at around 11 to 12 percent)
- The probability value corresponding to the false positive rate of 5% is 0.72.

THE NEW PICTURE
UNIVERSITY VS. ACADEMIC PATENTS

THE NEW PICTURE
UNIVERSITY VS. ACADEMIC PATENTS (2)

PUBLIC AND PRIVATE APPLICANT

NATIONAL AND FOREIGN APPLICANT

NATIONAL AND FOREIGN APPLICANT (2)

REGIONAL APPLICANT

NEXT STEPS