

Analytical report: Patterns of Science-Technology linkage

Science unfolding in technology: A cross-country analysis of scientific citations in patents

November 2013

Authors:

Julie Callaert

Jan-Bart Vervenne

Bart Van Looy

Technical Contributions :

Tom Magerman

Xiaoyan Song

Wouter Jeuris

Special acknowledgment: Bart Thijs & Wolfgang Glanzel (ECOOM, KU Leuven)



Content

1. Background and research objective

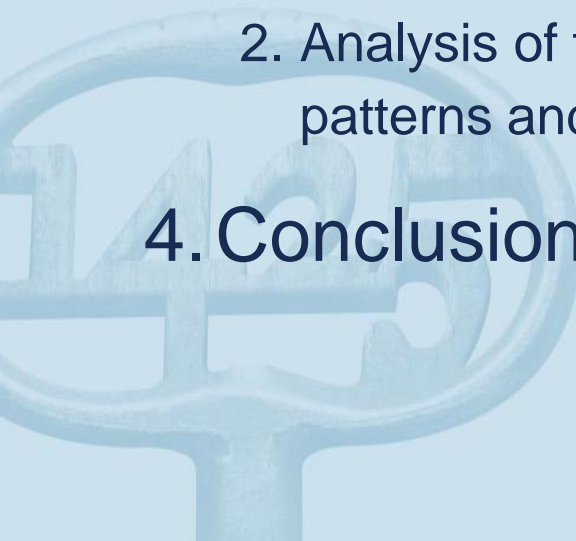
2. Data and methodology

3. Analyses and results

1. Mapping of cross-country citation flows between science and technology

2. Analysis of the relation between science-technology linkage patterns and national technological performance

4. Conclusions



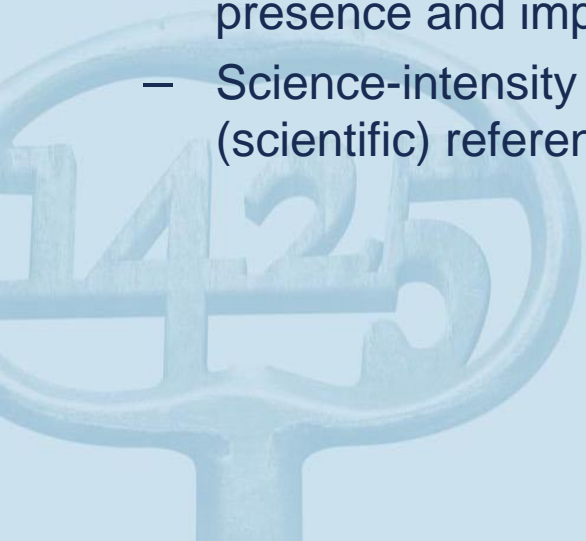
Background & Research objective

- **INTRODUCTION: SETTING THE STAGE**

- Importance of mapping and monitoring science-technology linkage patterns
- Relevance of patent-related indicators for measuring science-technology linkage
- Types of indicators: academic patenting; patent and non-patent citation indicators

- **OVERVIEW EXISTING EVIDENCE**

- Academic footprints in the technological landscape: evidence on the presence and impact of academic patents
- Science-intensity of technology: presence and meaning of non-patent (scientific) references in patent documents



Background & Research objective

- **RESEARCH OBJECTIVE**
 - Large scale mapping of cross-country citation flows (matrices) between science and technology: Patents citing scientific literature
 - Analyzing country-level and cross-country ST linkage patterns
 - Do supply and ‘absorption’ of science in technology coincide on a national level?
 - Mapping of cross-country citation flows between science and technology?
 - Do these patterns relate to technological performance (on the level of national innovation systems)?

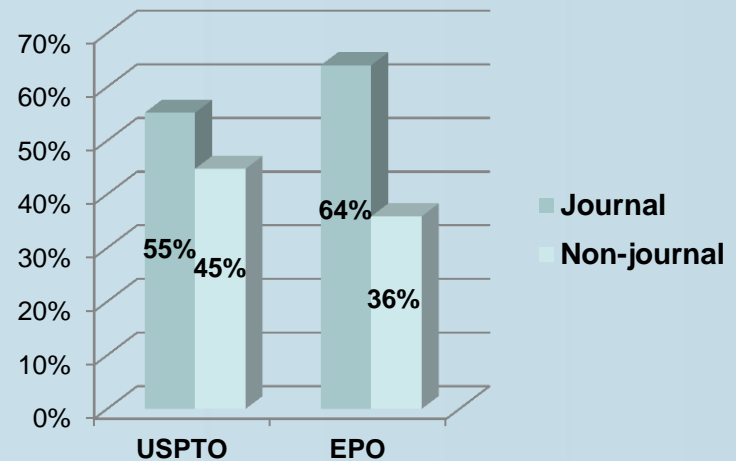
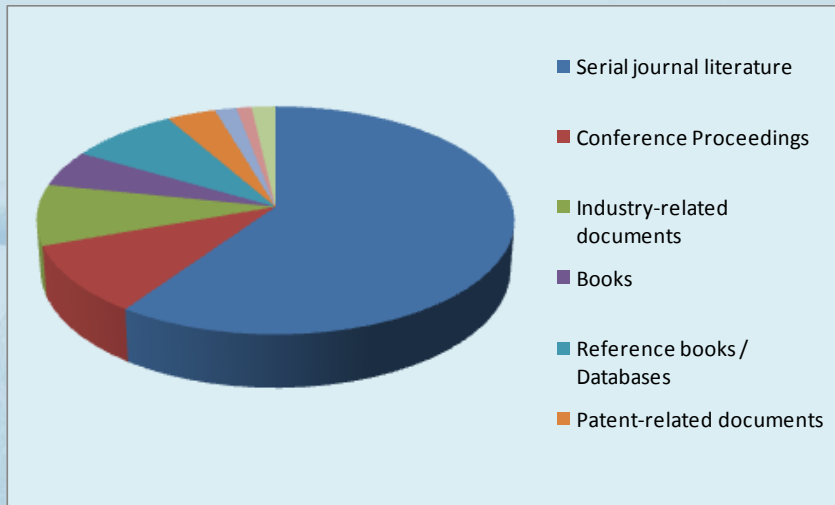


Data & Methodology

- Matching non-patent references to the Web of Science: development of text-mining based methodology
- Patents < PATSTAT database
 - EPO, USPTO and WIPO corporate patents, APPY 2000-2009
 - Applicant countries (citing)
- Scientific publications < Web of Science & Proceedings database (ISI/Thompson Reuters)
 - Journal papers and proceedings from 1990 onwards
 - Author countries (cited)
- Global coverage on citing and cited side: OECD member countries, EU-27 (and EFTA) member states, EU candidate countries, BRIC countries and Taiwan.

Identification of scientific NPRs

- Measuring and monitoring science-technology relations: indicators based on non-patent references ('science') in patent documents ('technology')
- But: many non-patent references are not strictly 'scientific' e.g.
 - *Michael Segell, Sports Illustrated, 1985, 1 pg.*
 - *License Agreement Between Dr. Albert M. Kligman and Johnson & Johnson, Jul. 18, 1984*
 - *USCU Sales Brochure 6-74/5070107.*



Source: Callaert, J., Van Looy, B., Verbeek, A., Debackere, K. and B. Thijs. (2006), *Traces of Prior Art: An analysis of non-patent references found in patent documents*. *Scientometrics*, 69(1), 3-20.

Identification of scientific NPRs

- Relevance of identifying scientific (i.e. peer-reviewed serial journal literature and proceedings) versus ‘non-scientific’ NPRs
- Development of a supervised machine learning approach:
 - Training set of 25,000 NPRs: classified by human raters
 - Document-by-term matrix → Linear Discriminant Analysis → assigning coefficients to terms

Illustrative Example: "Abstract of workshop"

$$\text{DOCUMENT} = \text{constante} + \text{coefficient} * (\# \text{ term frequency} - \text{mean correction}) + \text{coefficient} * (\# \text{ term frequency} - \text{mean correction})$$

$$\text{DOCUMENT} = 1,3014 + -4,4245 * (\# \text{ frequency of term 'abstract' } - 0,0021) + 5,9842 * (\# \text{ frequency of term 'workshop' } - 0,004)$$

$$\text{DOCUMENT} = 1,3014 + -4,4245 * 0,9979 + 5,9842 * 0,996$$

$$\text{DOCUMENT} = 2,8578$$

hence: scientific

Constant	1,3014	
Term	Coefficient	Mean Correction
workshop	5,9842	0,0044
symposium	5,1527	0,0114
meet	4,4552	0,008
abst	-4,4254	0,0021
disclosur	-4,2663	0,0113
confer	4,1981	0,0341
pna	4,1401	0,0022
genbank	-3,9761	0,0033
chapter	-3,7306	0,0055
handbook	-3,6599	0,0054
natur	3,5998	0,0167
transact	3,5579	0,0186
ed	-3,5447	0,0155
manual	-3,4234	0,0046
catalog	-3,4091	0,0047
acta	3,4032	0,0074
journal	3,3897	0,0759
proc	3,2414	0,0239
lett	3,1426	0,0073

Identification of scientific NPRs

- Application of classifier to all NPR's (N=14,826,883) in PATSTAT (version 10/2010) → +/- 50% were characterized as 'scientific'
- Currently:
 - Updatdng and validating the performance of the classifier (developed on PATSTAT version 10/2009; accuracy was 97%) on more recent PATSTAT versions)
- Resulting table (NPL_ID; scientific 0/1) will be made available to PATSTAT users.

Identification of scientific NPRs

Callaert, J., Grouwels, J., & Van Looy, B. (2011),
“Delineating the scientific footprint in technology:
Identifying science within non-patent references,”
Scientometrics, 91 (2), 383-398.

*Delineating the scientific footprint
in technology: Identifying scientific
publications within non-patent references*

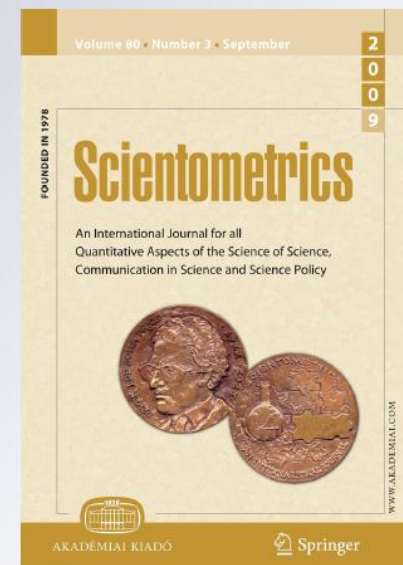
**Julie Callaert, Joris Grouwels & Bart Van
Looy**

Scientometrics

An International Journal for all
Quantitative Aspects of the Science of
Science, Communication in Science and
Science Policy

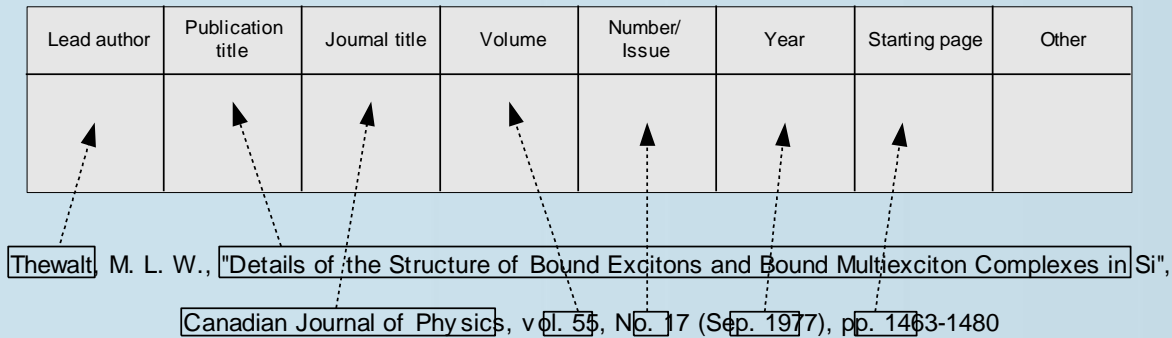
ISSN 0138-9130

Scientometrics
DOI 10.1007/s11192-011-0573-9



Matching scientific NPRs to WoS

- Previous work on linking NPRs to Web of Science (*Verbeek et al., 2002*) relied on parsing & matching approach



- But: no standard layout of NPR field in PATSTAT → making it difficult to reliably identify and parse subfields → loss of important information → matching based on incomplete data.
- After experimentation with different text-mining based techniques, a field-based matching approach was opted for: every (pre-)parsed field from a WoS article is matched separately to the set of NPR strings. A match is defined if sufficient fields from a WoS article are retrieved in an NPR text string.

Matching scientific NPRs to WoS

- Step 1: Calculation of field-based match scores for all NPR-WoS pairs within a priori defined year-based filter (+/- 2 years)
- Step 2: Download of NPR-WoS pairs with match scores above specified thresholds (*elimination of certain non-matches*)
- Step 3: Exploration and validation of additional filters on match scores to detect 'correct' matches (*precision: identification of certain matches*)
- Step 4: Extraction of the 'correct' NPR-WoS pairs, according to the criteria specified in step 3.

Matching scientific NPRs to WoS

- **Step 1: Calculation of field-based match scores for all NPR-WoS pairs within a priori defined year-based filter (+/- 2 years)**
 - 7 WoS fields: PY (publication year), AUTHOR_LN (last name of first author), SO (journal title), VL (volume), ISSUE (issue), BP (beginning page), TI (article title)
 - Match scores between a WoS-field and an NPR string are calculated as: *the number of distinct terms in the WOS field that occur also in the NPR text-string, divided by the number of distinct terms in the WOS field.*
 - Year filter: example

DERRYBERRY R.T. ET AL.: 'On CDMA 2000 evolution - Reverse high-speed packet data physical layer enhancements in CDMA 2000 1xEV-DV' IEEE COMMUNICATIONS MAGAZINE vol. 43, no. 4, 01 April 2005, pages 41 - 47, XP001228792

 - matched only to the subsets of WoS documents with a publication year between 1998-2002 and with a publication year between 2003-2007.
 - This a priori publication year filter reduces the amount of WoS documents to be considered for every NPR by a factor 6.

Matching scientific NPRs to WoS

- ***Step 2: Download of NPR-WoS pairs with match scores above specified thresholds (elimination of certain non-matches)***
 - Two a priori filters to eliminate certain non-matches (reducing the amount of data to be processed - only storing match scores for potentially correct matches).
 - Definition of filters: based on the fact that it requires multiple field matches to result in an overall NPR-WoS match, and that the value of some fields is more important in the matching process compared to other fields (cf. also Verbeek et al., 2002).



Matching scientific NPRs to WoS

A PRIORI FILTER 1ST LAYER

****(Sum of scores for the fields PY, AUTHOR_LN, VOLUME, ISSUE, BP >= 3.5)*

OR

*(Sum of scores for the fields PY, AUTHOR_LN, VOLUME, ISSUE, BP >= 3.0 **AND** at least 1 partial match on AUTHOR_LN (i.e. AUTHOR_LN > 0.5))*

OR

*(Sum of scores for the fields PY, AUTHOR_LN, VOLUME, ISSUE, BP >= 2.5 **AND** at least 1 substantial match on TI (TI >= 0.7 and TI contains at least 4 distinct terms))****

A PRIORI FILTER 2ND LAYER

****(A full or partial match on the following 5 fields: PY, AUTHOR_LN, VOLUME, ISSUE, BP)*

OR

(A full or partial match on 4 out of the following 5 fields: PY, AUTHOR_LN, VOLUME, ISSUE, BP

AND

(partial match on AUTHOR_LN (AUTHOR_LN >= 0.5))

OR

*(partial match on TI (TI >= 0.4) **AND** TI contains at least 3 distinct terms)*

OR

(partial match on SO (SO >= 0.6))

OR

(A full or partial match on 3 out of the following 5 fields: PY, AUTHOR_LN, VOLUME, ISSUE, BP

AND a partial match on AUTHOR_LN (AUTHOR_LN >= 0.6)

AND

*(a partial match on TI (TI >= 0.5) **AND** TI contains at least 3 distinct terms)*

OR

*(partial match on SO (SO >= 0.6))****

Matching scientific NPRs to WoS

- **Step 2: Download of NPR-WoS pairs with match scores above specified thresholds (elimination of certain non-matches)**
 - Filters resulting in a considerable reduction in the number of potentially matchable NPRs:

Application year	total # NPRs	# NPRs after initial 2 filters	% reduction
1993	352.816	46.426	87%
1994	460.503	85.839	81%
1995	723.368	157.986	78%
1996	532.667	142.728	73%
1997	674.315	199.290	70%
1998	703.119	232.285	67%
1999	812.603	292.806	64%
2000	965.667	371.207	62%
2001	1.027.899	414.389	60%
2002	1.107.040	479.656	57%
2003	1.066.291	462.135	57%
2004	1.105.543	486.554	56%
2005	1.026.694	452.833	56%
2006	836.612	367.681	56%
2007	614.495	271.395	56%
2008	378.668	181.931	52%
2009	115.982	76.809	34%



Matching scientific NPRs to WoS

- ***Step 3: Exploration and validation of additional filters on match scores to detect 'correct' matches (precision: identification of certain matches)***
 - A priori filter from step 2: sufficiently low to assure recall for matchable NPRs
 - Resulting matches with remaining NPRs: still considerable proportion of false positives
 - Need for identifying threshold match score values beyond which resulting pairs are certain matches (eliminating false positives and assuring precision).
 - Exploration of several filters (different combinations of match scores for considered fields): human rater validation of random sets
 - Decision criteria for withholding filters:
 - volume of retrieved matches
 - precision of matching results

Matching scientific NPRs to WoS

FILTER 1 MATCH SCORE TITLE $\geq 0,9$

FILTER 2 (excluding matches from FILTER1)

MATCH SCORE AUTHOR = 1

AND

MATCH SCORE TITLE $\geq 0,7$

FILTER 3 (excluding matches from FILTERS 1 and 2)

MATCH SCORE SOURCE $\geq 0,9$

AND

MATCH SCORE AUTHOR = 1

AND

MATCH SCORE VOLUME = 1

AND

MATCH SCORE BEGINNING PAGE = 1

FILTER 4 (excluding matches from FILTERS 1, 2 and 3)

MATCH SCORE AUTHOR = 1

AND

MATCH SCORE VOLUME = 1

AND

MATCH SCORE BEGINNING PAGE = 1

AND

LENGTH OF NPR TEXT STRING < 86

AND

CLASS NPR = 1



Matching scientific NPRs to WoS

- **Step 3: Exploration and validation of additional filters on match scores to detect 'correct' matches (precision: identification of certain matches)**

- Obtained accuracy (based on human rater validation of random subsets of matched pairs): 97%
- Obtained recall :

Appl year	# unique NPRs	% unique NPRs matched in FILTER 1	% unique NPRs matched in FILTER 2	% unique NPRs matched in FILTER 3	% unique NPRs matched in FILTER 4	TOTAL % MATCHED UNIQUE NPRs
1993	46.426	46%	9%	7%	7%	68%
1994	85.839	50%	10%	7%	8%	74%
1995	157.986	52%	10%	8%	9%	77%
1996	142.728	53%	9%	6%	7%	74%
1997	199.290	54%	9%	6%	6%	74%
1998	232.285	55%	9%	5%	6%	74%
1999	292.806	56%	8%	5%	6%	74%
2000	371.207	55%	8%	5%	6%	72%
2001	414.389	55%	8%	5%	5%	70%
2002	479.656	56%	7%	4%	4%	71%
2003	462.135	54%	7%	4%	4%	68%
2004	486.554	53%	6%	4%	4%	66%
2005	452.833	53%	6%	4%	4%	66%
2006	367.681	54%	6%	3%	3%	66%
2007	271.395	54%	6%	4%	4%	67%
2008	181.931	55%	5%	4%	6%	69%
2009	76.809	61%	5%	2%	2%	70%



Matching scientific NPRs to WoS

- **Step 4: Extraction of the 'correct' NPR-WoS pairs, according to the criteria specified in step 3.**

APPL YEAR	MATCHABLE NPRs AFTER STEP 2	MATCHED NPRs AFTER STEP 3	% MATCHED NPRs (MATCHABLE ONES)
1993	46.426	31.738	68%
1994	85.839	63.445	74%
1995	157.986	121.589	77%
1996	142.728	105.283	74%
1997	199.290	148.210	74%
1998	232.285	172.602	74%
1999	292.806	216.936	74%
2000	371.207	266.506	72%
2001	414.389	291.999	70%
2002	479656	340.019	71%
2003	462.135	314.649	68%
2004	486.554	319.131	66%
2005	452.833	298.568	66%
2006	367.681	241.328	66%
2007	271.395	181.002	67%
2008	181.931	125.196	69%
2009	76.809	53.591	70%

Matching scientific NPRs to WoS

- **Step 4: Extraction of the 'correct' NPR-WoS pairs, according to the criteria specified in step 3.**

APPL YEAR	TOTAL # NPRs	MATCHABLE NPRs AFTER STEP 2	MATCHED NPRs AFTER STEP 3	% MATCHED NPRs (/TOTAL)	% MATCHED NPRs (/MATCHABLE ONES)
1993	352.816	46.426	31.738	9%	68%
1994	460.503	85.839	63.445	14%	74%
1995	723.368	157.986	121.589	17%	77%
1996	532.667	142.728	105.283	20%	74%
1997	674.315	199.290	148.210	22%	74%
1998	703.119	232.285	172.602	25%	74%
1999	812.603	292.806	216.936	27%	74%
2000	965.667	371.207	266.506	28%	72%
2001	1.027.899	414.389	291.999	28%	70%
2002	1.107.040	479.656	340.019	31%	71%
2003	1.066.291	462.135	314.649	30%	68%
2004	1.105.543	486.554	319.131	29%	66%
2005	1.026.694	452.833	298.568	29%	66%
2006	836.612	367.681	241.328	29%	66%
2007	614.495	271.395	181.002	29%	67%
2008	378.668	181.931	125.196	33%	69%
2009	115.982	76.809	53.591	46%	70%

Analyses & results

Cross-country ST citation patterns (patents citing scientific literature)

→ main questions relate to the presence of a home bias or of geographic proximity effects in the citations from a country's corporate patents towards scientific resources

- Citation matrices: mapping countries of citing corporate patents (rows) and countries of origin of the cited scientific references (columns).
- Cell values represent relative intensities of citation linkages between the citing countries (rows) and the cited countries (columns):

$$CCC_{ij} = \frac{\text{citations of corporate citing country } i \text{ to science cited country } j}{\text{total citations of corporate citing country } i} \bigg/ \frac{\text{citations to science cited country } j}{\text{total corporate citations}}$$

- Global coverage: OECD member countries, EU-27 (and EFTA) member states, EU candidate countries, BRIC countries and Taiwan (on citing and cited side).

Analyses & results

Observations:

- Home 'bias': Countries cite their own science more intensively than science from foreign countries
- Extent of home 'bias' differs between countries:
 - countries where the within-country citation intensity is lowest are US, UK, DE (index below 2) and JP, FR, CH, CA, NL (index below 3) → countries that rank high on STI indicators
- Geographic proximity seems to matter
 - Examples of groups of neighboring countries that display relatively strong citation linkage intensities include: Austria, Switzerland, Germany – Denmark, Norway, Sweden – Sweden, Finland – France, Belgium.

Analyses & results

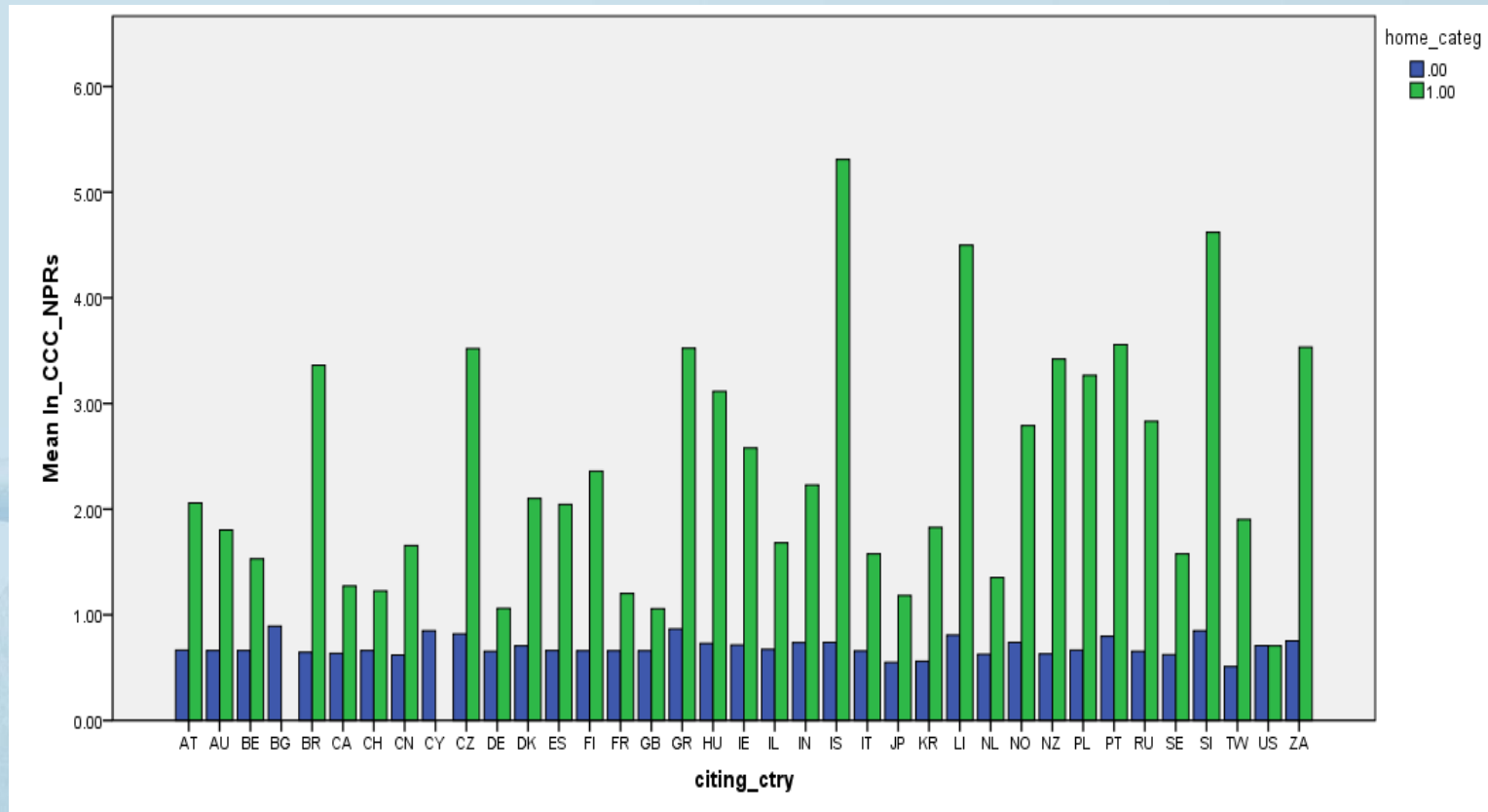
Statistical validation:

ANCOVA - Tests of Between-Subjects Effects						
Dependent Variable: Relative citation intensity between citing and cited country						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Parameter Estimate (B)
Corrected Model	3004,044 ^a	136	22,089	116,576	,000	
Intercept	301,668	1	301,668	1592,107	,000	2,738
Home (1/0)	358,625	1	358,625	1892,709	,000	1,530
Geographic distance between citing and cited country	17,118	1	17,118	90,343	,000	-,046
Citing country * Home	217,972	32	6,812	35,950	,000	
Technology Domain (FhG35)	614,326	34	18,068	95,359	,000	
Citing Country	840,003	34	24,706	130,390	,000	
Cited Country	556,623	34	16,371	86,402	,000	
Error	3219,408	16991	,189			
Total	19716,133	17128				
Corrected Total	6223,452	17127				

a. R Squared = ,483 (Adjusted R Squared = ,479)

Analyses & results

Interaction home 'bias' * citing country



Analyses & results (3.3)

3.3. Relation between characteristics of ST linkage patterns and national technological performance

- Dependent variable: National technological performance = citing country's patent volume per capita
- Independent variables: National-level characteristics of ST linkage patterns (based on the analyses reported in the previous sections):
 - strength of a country's home advantage in citing science = ratio between 'within country citation intensity' (CCC_home) and the average citation intensity between the source country and cited foreign countries (CCC_foreign)
 - concentration of foreign citations over cited countries = Herfindahl index is the sum of the squares of the (received) citations shares of all cited foreign countries, where the citation shares are expressed as fractions of the total number of citations to foreign countries → lower Herfindahl index indicates a geographically broader pattern of foreign citations
- Control variables: technology domain (FhG35 classification); national scientific performance citing country (per capita); science intensity of national patent portfolio; academic patenting.

Analyses & results

ANCOVA - Tests of Between-Subjects Effects						
Dependent Variable: Number of Patents per mio capita (Citing Country)						
Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Parameter Estimate (B)
Corrected Model	2282,367 ^a	39	58,522	100,608	,000	
Intercept	2,230	1	2,230	3,834	,050	1,221
Scientific performance (Citing Country)	656,199	1	656,199	1128,096	,000	2,958
Academic patenting (Citing Country)	20,438	1	20,438	35,136	,000	,245
Science intensity of patent portfolio (Citing Country)	79,180	1	79,180	136,121	,000	,218
Herfindahl Foreign Citations	2,710	1	2,710	4,658	,031	-,435
Ratio Home versus Foreign citation intensity	,091	1	,091	,156	,693	-,013
Technology domain (FhG35)	298,384	34	8,776	15,087	,000	
Error	566,563	974	,582			
Total	19157,226	1014				
Corrected Total	2848,930	1013				

a. R Squared = ,801 (Adjusted R Squared = ,793)

Preliminary conclusions

- Sufficient levels of science-technology linkage are imperative for the performance of national innovation systems.
- As such our findings underscore the relevance of science (for technology)
- Both scientific publications, the science intensity of patents and even the patenting behavior of academia are positively related with technical performance.
- On top, countries which 'source' science from a variety of countries, seem to benefit from this diversity.
- Further issues : causality (both ways) - field differences (implying a breakdown of scientific indicators as well) ...

Thank you !

Questions?

Suggestions/Recommendations?



Analyses & results

ANCOVA - Tests of Between-Subjects Effects → SPLIT FILE BASED ON SCIENTIFIC PRODUCTIVITY

Dependent Variable: Number of Patents per mio capita (Citing Country)

Source	LOW SCIENTIFIC PRODUCTIVITY						HIGH SCIENTIFIC PRODUCTIVITY					
	Type III Sum of Squares	df	Mean Square	F	Sig.	Parameter Estimate (B)	Type III Sum of Squares	df	Mean Square	F	Sig.	Parameter Estimate (B)
Corrected Model	1184,370 ^a	39	30,368	51,652	,000		283,693 ^a	39	7,274	22,809	,000	
Intercept	,500	1	,500	,850	,357	1,132	7,146	1	7,146	22,408	,000	1,971
Scientific Output per 1000 Capita (Citing Country)	234,845	1	234,845	399,438	,000	3,702	53,635	1	53,635	168,178	,000	2,645
Number of university patents per mio capita (Citing Country)	45,655	1	45,655	77,653	,000	,773	,384	1	,384	1,204	,273	,043
Number of SNPRs (Citing Country)	20,956	1	20,956	35,644	,000	,180	14,160	1	14,160	44,399	,000	,135
Ratio Home versus Foreign citation intensity	3,277	1	3,277	5,574	,019	-,102	,001	1	,001	,003	,958	,002
Herfindahl Foreign Citations	3,327	1	3,327	5,659	,018	-,619	,074	1	,074	,232	,630	-,134
Technology domain (FhG35)	149,561	34	4,399	7,482	,000		159,305	34	4,685	14,692	,000	
Error	267,513	455	,588				147,978	464	,319			
Total	6009,784	495					12961,258	504				
Corrected Total	1451,882	494					431,671	503				
	a. R Squared = ,816 (Adjusted R Squared = ,800)						a. R Squared = ,657 (Adjusted R Squared = ,628)					