Geographical and Lingual Preferences in Scientific Collaboration of the European Union (1994-2003)

LIMING LIANG[1,2]
LIN ZHANG[1]
HILDRUN KRETSCHMER[2,3]
ANDREA SCHARNHORST[4]

1 Institute for Science Technology and Society, Henan Normal University, Xinxiang, 453007 (China), pllm@public.xxptt.ha.cn

2 The School of Humanities and Social Sciences, Dalian University of Technology, Dalian, (China)

3 Department of Library and Information Science, Humboldt-University of Berlin, Dorotheenstr. 26, D-10099 Berlin (Germany)

4 The Royal Netherlands Academy of Arts and Sciences, Amsterdam (The Netherlands)

Abstract

In 2001 and 2004 we launched and relaunched a project to study the structure and mechanism of scientific collaboration among 15 member countries of European Union (EU-15) using bibliometric data (1994-2003). This paper presents the basic ideas, methods, and part of the findings. In scientific collaboration EU-15 gets more open over the past decade. The proportion of the collaboration within EU countries is decreasing, while the proportion of the collaboration of EU countries with also the co-authors from non-EU countries is increasing. There exist geographical and lingual preferences in EU-15 scientific collaboration. When the geographical proximity of the collaborative countries gets smaller, the collaborative strength negative-exponentially declines. When the lingual proximity of the collaborative countries gets smaller, the collaborative strength negative-exponentially declines, as well.

1. Introduction

As the largest political, economic union, the European Union has got much attention in different analysis from different points of view. One hot topic is the development of science and the emergence of a European Research Area. Since the beginning of 1990's different scientists have studied the scientific collaboration inside of European Community or European Union [1] [2] [3]. In 2001 we launched a new project to study the structure and mechanism of scientific collaboration among 15 member countries of European Union (EU-15 for short) using bibliometric data. In statistics of the publication records extracted from the Science Citation Index® we met a difficult problem. Some records contain as country name for example "England" in place of the UK. So, the field of country names contains also names of organization or regions or places. The question merges how to identify and clear such records? At that time we had only access to the

SCI CD-ROM version and we were not able and had not enough time to solve this problem. Since 2004 we have had the chance to use ISI's Web of Science®. The database's powerful function of retrieval and process help us to overcome the old problem. Therefore, we decided to restart the study in 2004.

2. Data and Methods

The data we used in this paper were searched during the period of January to February in year 2005. We used as database ISI's Web of Science, in the time span from 1994 to 2003. We restricted the search to the document type "article" and searched for "all language". The restriction of our sample was that we downloaded only articles with authors from at least two EU countries, no matter how many authors each EU country has, and no matter whether there is the author from non-EU countries. Such articles are denoted as EUC. The number of eligible records was 250.051.

Table 1 shows the distributions of the 250.051 EUC records in terms of the number of co-authored EU countries and the restriction of with or without non-EU country. Here, the type I-EUC represents the number of collaborative articles with authors from EU countries, without any author from a non-EU country. We call these articles "**I**nternal collaboration articles". The type E-EUC represents the co-authored article of EU countries, with at least one author from non-EU country ("**E**xternal collaboration articles"). The type T-EUC denotes all co-authored article of EU countries, no matter whether there is the co-author from non-EU country. It is the sum of I-EUC and E-EUC. The numbers in the title line of the table indicate how may EU countries are involved in the collaboration.

Table 1: Distribution of co-authored articles of 15 EU countries (1994-2003)

Publication	Type	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
1994	I-EUC	10160	659	101	24	15	11	5	0	1	0	0	0	0	0	10976
1994	E-EUC	3056	451	155	42	51	18	23	5	3	1	0	14	2	0	3821
1994	**T-EUC**	**13216**	**1110**	**256**	**66**	**66**	**29**	**28**	**5**	**4**	**1**	**0**	**14**	**2**	**0**	**14797**
1995	I-EUC	11193	810	87	35	12	9	2	1	0	0	0	0	0	0	12149
1995	E-EUC	3526	527	182	94	50	16	26	7	9	1	0	26	2	0	4466
1995	**T-EUC**	**14719**	**1337**	**269**	**129**	**62**	**25**	**28**	**8**	**9**	**1**	**0**	**26**	**2**	**0**	**16615**
1996	I-EUC	13440	972	125	50	14	13	3	2	2	0	0	0	0	0	14621
1996	E-EUC	4403	660	208	109	77	23	40	2	3	4	2	27	2	0	5560
1996	**T-EUC**	**17843**	**1632**	**333**	**159**	**91**	**36**	**43**	**4**	**5**	**4**	**2**	**27**	**2**	**0**	**20181**
1997	I-EUC	14633	1197	157	48	26	13	4	1	0	0	0	0	0	0	16079
1997	E-EUC	4910	778	245	112	97	27	41	6	3	4	5	18	0	0	6246
1997	**T-EUC**	**19543**	**1975**	**402**	**160**	**123**	**40**	**45**	**7**	**3**	**4**	**5**	**18**	**0**	**0**	**22325**
1998	I-EUC	16130	1413	211	54	23	16	0	1	0	0	1	1	0	0	17850
1998	E-EUC	5644	905	262	135	96	32	57	11	5	1	0	19	0	0	7167
1998	**T-EUC**	**21774**	**2318**	**473**	**189**	**119**	**48**	**57**	**12**	**5**	**1**	**1**	**20**	**0**	**0**	**25017**
1999	I-EUC	17239	1516	216	69	44	17	4	1	5	0	1	0	1	0	19113
1999	E-EUC	6443	989	282	150	101	28	36	7	4	2	1	28	0	0	8071
1999	**T-EUC**	**23682**	**2505**	**498**	**219**	**145**	**45**	**40**	**8**	**9**	**2**	**2**	**28**	**1**	**0**	**27184**
2000	I-EUC	17799	1694	199	89	31	25	9	1	1	2	0	1	0	0	19851
2000	E-EUC	6957	1187	339	169	99	30	42	12	1	2	1	33	0	0	8872
2000	**T-EUC**	**24756**	**2881**	**538**	**258**	**130**	**55**	**51**	**13**	**2**	**4**	**1**	**34**	**0**	**0**	**28723**
2001	I-EUC	18277	1863	276	84	33	25	5	1	0	0	0	0	0	0	20564

2001	E-EUC	7454	1311	388	207	94	51	28	10	2	4	12	6	0	1	9568
2001	**T-EUC**	**25731**	**3174**	**664**	**291**	**127**	**76**	**33**	**11**	**2**	**4**	**12**	**6**	**0**	**1**	**30132**
2002	I-EUC	18995	1920	262	99	29	15	6	2	2	1	0	1	0	1	21333
2002	E-EUC	8144	1334	464	199	107	59	22	40	4	3	6	0	0	0	10382
2002	**T-EUC**	**27139**	**3254**	**726**	**298**	**136**	**74**	**28**	**42**	**6**	**4**	**6**	**1**	**0**	**1**	**31715**
2003	I-EUC	19802	2077	297	81	46	14	2	3	0	0	0	0	0	1	22323
2003	E-EUC	8532	1495	473	247	147	64	35	19	6	2	18	1	0	0	11039
2003	**T-EUC**	**28334**	**3572**	**770**	**328**	**193**	**78**	**37**	**22**	**6**	**2**	**18**	**1**	**0**	**1**	**33362**
94-03	I-EUC	157668	14121	1931	633	273	158	40	13	11	3	2	3	1	2	174859
94-03	E-EUC	59069	9637	2998	1464	919	348	350	119	40	24	45	172	6	1	75192
94-03	**T-EUC**	**216737**	**23758**	**4929**	**2097**	**1192**	**506**	**390**	**132**	**51**	**27**	**47**	**175**	**7**	**3**	**250051**

To measure the collaborative strength of research collaboration we can use as many other authors the Salton formula[4]. However, in this study we use an Modified Salton formula to calculate the collaborative strength, or as we call it, Modified Salton index, between any two EU countries[5]:

$$s_{ij}=n_{ij}/(c_i c_j)^{1/2} \quad (i \neq j) \qquad (1).$$

Here, n_{ij} is the number of co-authored articles of country i and country j. c_i is the total number of co-authored articles of country i with other countries. c_j is the total number of co-authored articles of country j with other countries. The calculation of s_{ij} depends on the sample set. It means that n_{ij}, c_i and c_j should be determined based on a certain sample set, such as the set of all I-EUC articles, the set of all E-EUC articles or the set of all T-EUC articles. Formula (1) is used to calculate the Modified Salton index of the collaboration of two countries. Theoretically, it could be generalized to the collaboration of n countries. Especially, when $n = 3$, we obtain formula (2).

$$\mathbf{s_{ijk}=n_{ijk}/(c_i c_j c_k)^{1/3} \quad (i \neq j \neq k)} \qquad \textbf{(2)}$$

3. Results

3.1. EU countries have been active to extend the collaboration with non-EU countries

Analyzing the data in Table 1 we found three characters. First, from 1994 to 2003 the absolute values of I-EUC, E-EUC and T-EUC are all increasing (see Figure 1). Second, in T-EUC the share of I-EUC linearly declined and the share of E-EUC mounted up linearly (see Figure 2 and Figure 3), which made the ratio of I-EUC to E-EUC go down continuously (see Figure 4). In 1994 the ratio of I-EUC to E-EUC is as high as 2.87. After ten years, however, in 2003, the ratio is only 2.02, a drop of 85 percent. This illustrates that in international scientific collaboration the European Union gets more open towards non-EU countries. In the viewpoint of geography EU countries have been very active during the period of 1994-2003 to extend the collaborative scope and to seek more cooperators from non-EU countries. Third, the collaborative scale of E-EUC is larger than that of I-EUC. This phenomenon is easy to understand, because the authors of an E-EUC article come from not only the EU countries, but the non-EU country(ies).

Fig.1: Increasing trends of I-EUC, E-EUC and T-EUC (1994-2003)

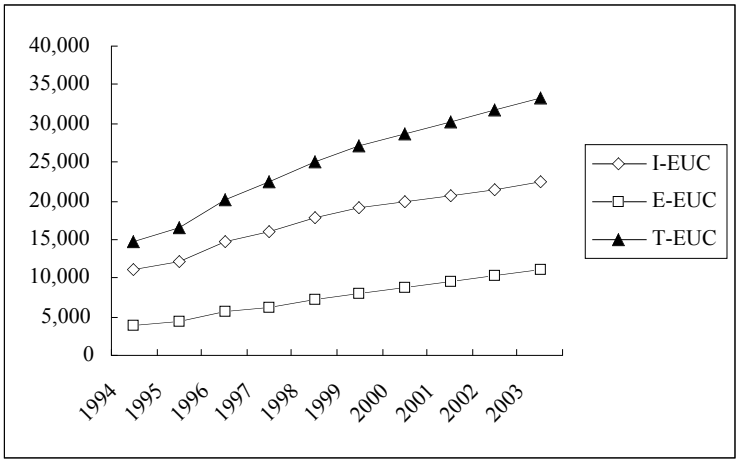


Fig. 2: Share of I-EUC in T-EUC

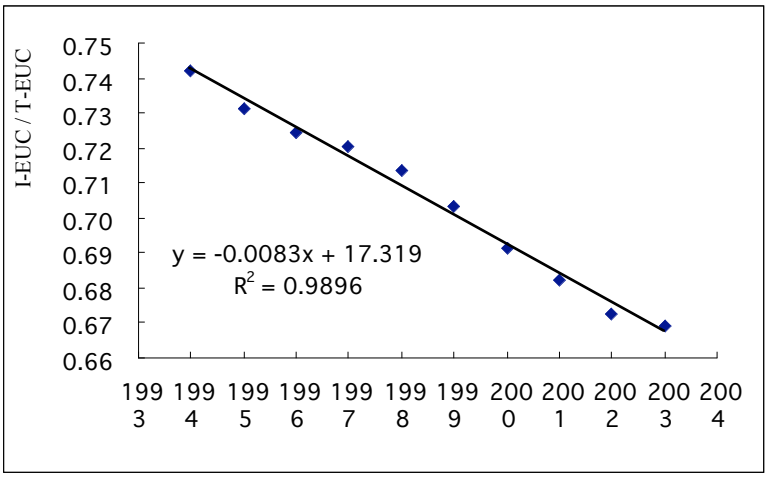


Fig. 3: Share of E-EUC in T-EUC

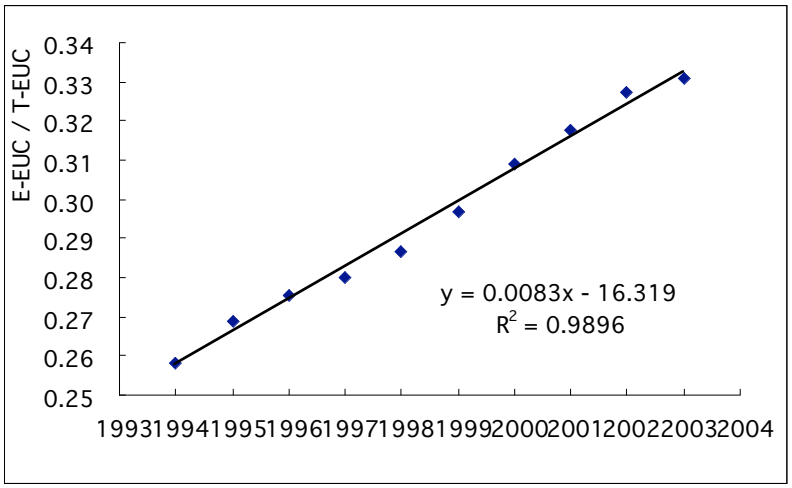
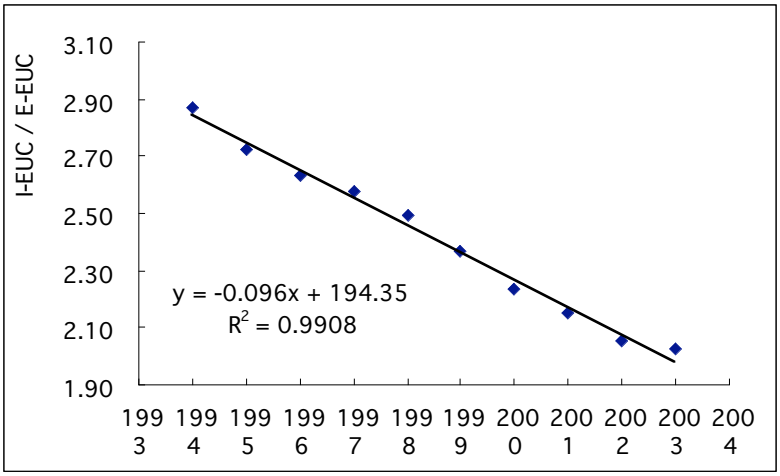


Fig. 4: The ratio of I-EUC to E-EUC



3.2. Geographical preference in scientific collaboration of 15 EU countries

We discuss this issue based on two sub-sets of I-EUC sample set: P2 and P3.

P2 article: Article co-authored by any couple of EU countries (no matter how many authors each country has), without the author from the third EU countries and non-EU countries. We call such article produced by and only by two EU countries P2 article. From Table 1 we know that the number of P2 records is 157,668.

P3 article: Article co-authored by any triple of EU countries (no matter how many authors each country has), without the author from the fourth EU countries and non-EU countries. We call such article produced by and only by three EU countries P3 article. From Table 1 we know that the number of P3 records is 14,121.

Table 2 is the matrix of P2 articles co-authored by any two EU countries.

Table 2: Matrix of P2 articles (1994-2003)

_	UK	GER	FR	IT	ES	NL	SE	BE	DK	AT	FI	PT	GR	IE	LU	C_i
UK		11705	9291	7377	5664	5470	3354	2269	2240	894	1328	1719	2048	2186	12	55557
GER	11705		10121	5967	3648	5706	3127	2323	2023	5484	1330	639	1312	432	63	53880
FR	9291	10121		7980	6100	2530	1588	5025	860	759	595	1229	1175	333	76	47662
IT	7377	5967	7980		3461	2197	1172	1117	664	852	371	354	488	273	6	32279
ES	5664	3648	6100	3461		1235	763	973	477	414	358	1104	217	167	6	24587
NL	5470	5706	2530	2197	1235		1070	2983	726	471	492	380	246	219	8	23733
SE	3354	3127	1588	1172	763	1070		530	2253	306	2213	211	148	116	4	16855
BE	2269	2323	5025	1117	973	2983	530		240	175	159	168	226	116	85	16389
DK	2240	2023	860	664	477	726	2253	240		166	389	120	93	63	2	10316
AT	894	5484	759	852	414	471	306	175	166		154	74	70	44	2	9865
FI	1328	1330	595	371	358	492	2213	159	389	154		53	45	78	2	7567
PT	1719	639	1229	354	1104	380	211	168	120	74	53		33	52	0	6136
GR	2048	1312	1175	488	217	246	148	226	93	70	45	33		31	1	6133
IE	2186	432	333	273	167	219	116	116	63	44	78	52	31		0	4110
LU	12	63	76	6	6	8	4	85	2	2	2	0	1	0		267
C_j	55557	53880	47662	32279	24587	23733	16855	16389	10316	9865	7567	6136	6133	4110	267	315336

Table 3 lists the Modified Salton indexes, calculated by formula (1), of the collaboration between any two EU countries. By formula (2) we also calculated the Modified Salton index of each triple.

Table 3: EU-15: Matrix of Salton index of P2 articles

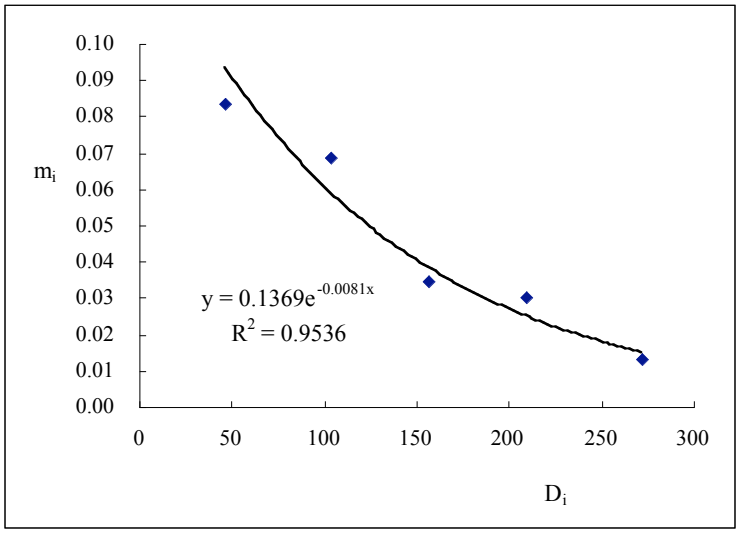
_	UK	GER	FR	IT	ES	NL	SE	BE	DK	FI	AT	PT	GR	IE	LU
UK		0.214	0.181	0.174	0.153	0.151	0.11	0.075	0.094	0.065	0.038	0.093	0.111	0.145	0.003
GER			0.2	0.143	0.1	0.16	0.104	0.078	0.086	0.066	0.238	0.035	0.072	0.029	0.017
FR				0.203	0.178	0.075	0.056	0.18	0.039	0.031	0.035	0.072	0.069	0.024	0.021
IT					0.123	0.079	0.05	0.049	0.036	0.024	0.048	0.025	0.035	0.024	0.002
ES						0.051	0.037	0.048	0.03	0.026	0.027	0.09	0.018	0.017	0.002
NL							0.053	0.151	0.046	0.037	0.031	0.031	0.02	0.022	0.003
SE								0.032	0.171	0.196	0.024	0.021	0.015	0.014	0.002
BE									0.018	0.014	0.014	0.017	0.023	0.014	0.041
DK										0.044	0.016	0.015	0.012	0.01	0.001
FI											0.018	0.008	0.007	0.014	0.001
AT												0.010	0.009	0.007	0.001
PT													0.005	0.010	0
GR														0.006	0.001
IE															0
LU	_	_	_	_	_	_	_	_	_	_	_	_	_	_	_

Based on the Modified Salton indexes of couples and triples in P2 and P3 sample sets we explore the geographical preference in scientific collaboration of 15 EU countries. Three methods are used. Method 1: measuring the geographical proximity of two countries by the “crow fly distance” between two capitals

Define the crow fly distance between two capitals as the distance between two countries. We measured the length of the line between any two capitals based on a map of the world (http://www.slammers.de/agentur-dateien/_europe%20map.htm). Among the 105 distances of any two EU countries, the longest is 304, the shortest is only 15.5, and their difference is 288.5. Divide 288.5 into 5 groups. They are: 15.5-73.2; 73.2-130.9; 130.9-188.6; 188.6-246.3 and 246.3-304. The five groups contain 28, 33, 26, 13 and 5 distance data, respectively. Or we say the five groups consist of 28, 33, 26, 13 and 5

couple of countries, respectively. Denote D_i as the mean distance of each group, m_i as the average Modified Salton index of each group (i=1, 2, 3, 4, 5). Figure 5 shows the relationship between D_i and m_i. It is a negative exponential distribution.

Fig. 5 Relationship between D_i and m_i (P2 sample set, method 1)

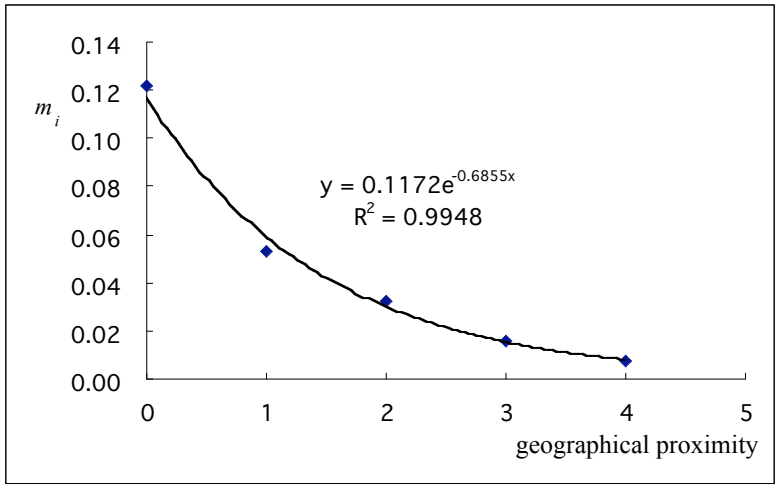


- Method 2: measuring the geographical proximity of two countries by the minimum number of steps going from one country to another

In the second method, the geographical proximity defined as the minimum number of passing countries when going from one country to another country over land/or taking a passage over land. For two adjoining countries the geographical proximity is defined as 0. If we can not reach country B from country A just via the land passage, we have to cross the channel, the sea or the ocean, the geographical proximity of A and B is determined as follows: the step(s) via continental country(ies) plus the step of the waterway, here, though a channel plus 0, through the sea or ocean plus 1.

.Finally, we determined the geographical proximity of any two countries. Among the 105 couple of countries, the numbers of couples with geographical proximity as 0_1_2_3 and 4 are 23_39_28_13 and 2, respectively. According to the values of geographical proximity, naturally, all the couples could be classified into 5 groups. The first group covered 23 couple of countries, and the last group contains only two couples. Figure 6 shows the m_i values of each group. The distribution of the 5 m_i values is a perfect negative exponential distribution.

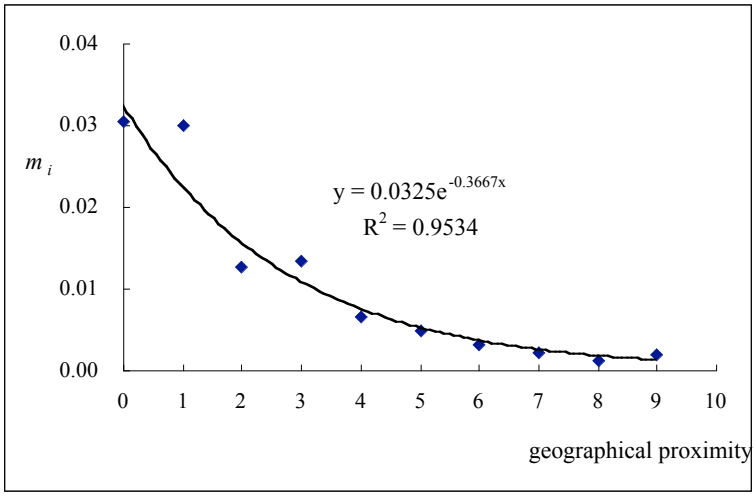
Fig. 6 Relationship between geographical proximity and m_i (P2 sample set, method 2)



Method 2 can also be used in analyzing of the relationship between geographical proximity and m_i based on P3 data. Theoretically, among 15 EU countries there are 455 different triples. Actually, during the period 1994-2003 only 339 triples published P3 articles. From Table 1 we found that the number of P3 records is 14,121.

How can we measure the geographical proximity of a triple of countries A, B and C? A simple method is to calculate the sum of the geographical proximities of A and B, B and C, A and C. Here, the geographical proximity of any two countries is defined as the minimum number of steps going from one country to another. The calculation results of the 339 triples show that the maximum sum is 9, the minimum is 0, i.e. the triple of three countries adjoining each other. In this way we get ten groups with the geographical proximities values 0, 1, 2, … , 9, respectively. Calculate m_i for each group. Figure 7 represents the relationship between geographical proximity and m_i. Again we found an exponential distribution.

Fig. 7 Relationship between geographical proximity and m_i (P3 sample set, method 2)



- Method 3: measuring the geographical proximity of two countries by judging

whether they are adjoining countries

The third method used to present the geographical preference in scientific collaboration between two EU-15 countries is a very simple judging: firstly to determine the closest collaborators of each country by the Modified Salton index, and then to count how many closest collaborators are the adjoining countries. The identified results related to P2 articles are listed in Table 4. We found that the proportion of adjoining regions being as the closest collaborators is as high as 80%. We also noticed that all the three exceptions (without star) refer to the two most productive countries: UK and Germany. It makes us recall the similar situation in the regional collaboration between China's regions, where the most productive region Beijing plays the same role [5].

Table 4: Two closest collaborator of each country in P2 publications

	Closest			Closest
Uk	Germany	_	Denmark	Sweden*
Germany	Austria*	_	Austria	Germany*
France	Italy*	_	Finland	Sweden*
Italy	France*	_	Portugal	Uk
Spain	France*	_	Greece	Uk
Netherlands	Germany*	_	Ireland	Uk*
Sweden	Finland*	_	Luxembourg	Belgium*
Belgium	France*	_		

Note: * denotes the adjoining country

3.3. Lingual preference in scientific collaboration of 15 EU countries

Communication takes place using languages. The same holds for scientific communication. In scientific collaboration, one could expect that if the collaborators share the same or similar language this would make the academic exchange easier. Is there the lingual preference in scientific collaboration among 15 EU countries? This is an interesting problem. In this paper we will try to find the answer based on the sample set of 157.668 P2 articles.

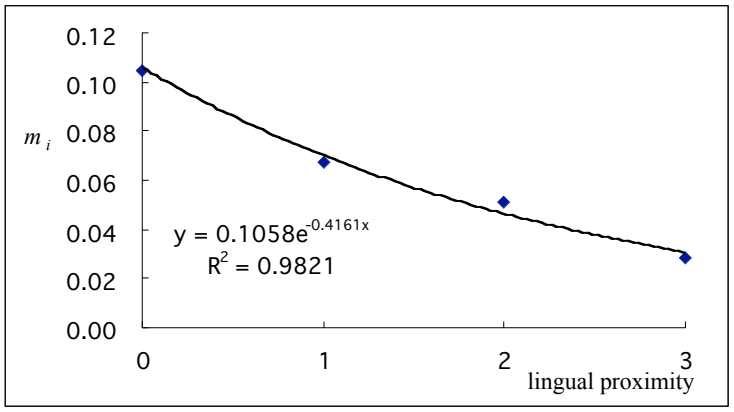
The method we used is not complicated. First, the information about the official language(s) of every country is needed. Second, the knowledge background of the languages' family is necessary. Table 5 lists the official language(s) of each country and the corresponding lingual group(s) and branch(es). Third, define the concept of "lingual proximity". To compare the official languages of two collaborative countries, if their official languages (or one of the official languages) are the same, we denote their lingual proximity as 0; if their official languages are different, but belong to the same lingual group and the same lingual branch, their lingual proximity is denoted a 1; if their official languages belong to the same group, but different branches, their lingual proximity is denoted as 2; if their official languages belong to different groups, their lingual proximity is denoted as 3. Fourth, put the couples of collaborative countries with the same lingual proximity in the same group, all the 105 couples are classified into 4 groups. The groups with the lingual proximity 0, 1, 2 and 3 have 8, 30, 53 and 14 couples, respectively. Finally, calculate the mean Modified Salton index m_i for each group and find the

relationship between Modified Salton indexes and the lingual proximity. Figure 8 shows a negative exponential distribution.

Table 5: Official languages and the corresponding language branches

Country	Official Language	Lingual Group	Lingual Branch
UK	English	Germanic Lingual Group	West-Germanic Lingual Branch
GER	German	Germanic Lingual Group	West-Germanic Lingual Branch
FR	French	Germanic Lingual Group	East-Germanic Lingual Branch
IT	Italian	Germanic Lingual Group	East-Germanic Lingual Branch
ES	Spanish	Germanic Lingual Group	East-Germanic Lingual Branch
NL	Dutch	Germanic Lingual Group	West-Germanic Lingual Branch
SE	Swedish	Germanic Lingual Group	North-Germanic Lingual Branch
BE	French / Flemish	Germanic Lingual Group	East-Germanic Lingual Branch / West-Germanic Lingual Branch
DK	Denish	Germanic Lingual Group	North-Germanic Lingual Branch
FI	Finnish / Swedish	Finno-Ugric Lingual Group /Germanic Lingual Group	North-Germanic Lingual Branch
AT	German	Germanic Lingual Group	West-Germanic Lingual Branch
GR	Greek	Greek Lingual Group	
PT	Portuguese	Germanic Lingual Group	East-Germanic Lingual Branch
IE	Irish / English	Cornish (Celtic) Lingual Group /Germanic Lingual Group	Goidelic Lingual Branch / West-Germanic Lingual Branch
LU	French / German /Luxembourgish	Germanic Lingual Group	East-Germanic Lingual Branch / West-Germanic Lingual Branch

Fig. 8 Relationship between lingual proximity and m_i (P2 sample set)

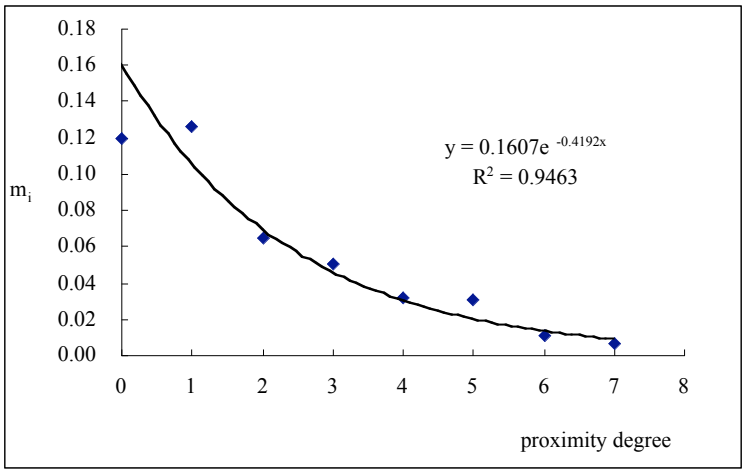


3.4. Associated effect of geographical preference and lingual preference in scientific collaboration of 15 EU countries

When analyzing the geographical preference or the lingual preference we found the negative exponential distribution between the mean Modified Salton index and the

geographical or lingual proximity. What is the associated effect of the geographical preference and the lingual preference?
We calculated the sum of the geographical proximity (based on method 2) and the lingual proximity for each couple of countries, which is called the proximity degree of the two countries. The 105 couples obtain 105 proximity degree, which fall into 8 groups with the proximity degree 0, 1, 2, 3, 4, 5, 6 and 7, respectively. Calculate the mean Modified Salton index m_i for each group. The relationship between the mean Modified Salton index and the proximity degree is shown in Figure 9. It is also the negative exponential distribution.

Fig. 9 Relationship between proximity degree and m_i



4. Conclusion and discussion

Some conclusions are drawn from the above analyses.

In scientific collaboration European Union gets more open over the past decade. The proportion of the collaboration within EU countries is decreasing. The proportion of the collaboration of EU countries with also the co-authors from non-EU countries is increasing. There exists geographical preference in scientific collaboration of the European Union. When the geographical proximity of the collaborative countries gets smaller, the collaborative strength negative-exponentially declines. The probability of adjoining country being as the closest collaborator is as high as 80%.

Lingual preference is one of the most important factors affecting research cooperation of the European Union. When the lingual proximity of the collaborative countries gets smaller, the collaborative strength negative-exponentially declines.

In our former study on the regional scientific collaboration of China and this study on EU countries we have found that country's or regional scientific productivity is another important factor affecting the collaborative strength. We will discuss this problem in the future.

1 References H.F. Moed_R.E.De Bruin_A.J. Nederhof and R.J.W .Tijssen_International scientific co-operation and awareness within the European Community: Problems and perspectives_*Scientometrics*_1991(21)_291-311

2 F. Narin_K.Stevens, Edrth S and Whitlow_Scientific co-operation in Europe and the citation of multinationally authored papers_*Scientometrics*_1991(21), 313-323

3 A. Schubert, W. Glanzel and H.-J. Czerwon_A bibliometric analysis of international scientific collaboration of the European Union (1985-1995)_*Scientometrics*_1999(45), 185-202

4 G.Salton and M.J.Mcgill_Introduction to modern information retrieval_McGraw-Hill_N.Y._1983

5 L. Liang and Zhu L_Major factors affecting China's inter-regional research collaboration:Regional scientific productivity and geographical proximity_*Scientometrics*_2002(55)_287-316