Computer scientists from the former USSR: international mobility patterns and scientific success

Andrey Indukaev
IDHE, ENS Cachan and EUSP, St. Petersburg, Russian Federation
indukaev.a@gmail.com

Andrei Mogoutov
Andrei Mogoutov, IFRIS, Paris, France

Vincent Lepinay
EUSP, St. Petersburg, Russian Federation and SciencesPo, Paris, France

ABSTRACT

In the present paper, we develop a new method of longitudinal analysis of bibliographic data in order to explore international mobility of researchers from the former USSR through their publication activity.

Firstly, by means of name recognition algorithm using machine learning, we extracted from Web of Science a dataset of publications of more than three thousand of the most active computer scientists from the former Soviet Union. Then, the information on individuals' scientific production is presented in the form of a sequence of states which summarizes the affiliation location for all articles published by a certain author in a given period.

We use Optimal Matching algorithm to measure the degree of difference (which, in the sequence analysis, is called distance) between the sequences of individual researchers' activity. The distance between sequences is analyzed by means of hierarchical clustering, which permits us to group computer scientists from the former USSR in several classes according to publication activity patterns.

Not surprisingly, ex-soviet researchers having permanent affiliation in their home country are cited less than those who have permanent foreign affiliation. However, those who switch affiliations from former USSR to foreign or the other way round and publish in internationalized groups have one of the highest levels of citation per article among newcomers in discipline.

Our research shows that scientific mobility of successful authors can be not only unidirectional, but can take form of a complex go-and-return pattern, the claim which relativizes the "brain drain" paradigm in the analysis of migration of highly qualified specialists from the former URSS. On the methodological level, we propose a new method for analyzing scientific activity which takes into account its longitudinal dynamics. This method can be used for research questions going far beyond the scope of migration studies.

Funding for this research was provided by the Mega-grant of the Russian Federation Government to support scientific research under the supervision of leading scientist at the European University at St.Petersburg, No.14.U04.31.0001.

CEE-SECR '14, October 23 - 25 2014, Moscow, Russian Federation

ACM 978-1-4503-2889-0/14/10$15.00

http://dx.doi.org/10.1145/2687233.2687257

Categories and Subject Descriptors

G.3 [Time series analysis]

General Terms

Management, Performance, Human Factors, Theory

Keywords

Optimal matching, sequence analysis, scientometrics, bibliometrics, academic career, migration, Russia, former USSR

1. INTRODUCTION

The migration of highly skilled professionals is an important issue for many sectors, especially for IT. It has already been noted that highly skilled labor markets are crucial for a wide variety of economic sectors. The now famous examples of the San Francisco Bay Area and the Boston cluster are mobilized regularly by innovation consultants and by regional planners seeking to replicate the miracle of university/industry collaborations. Around these areas, many economic sectors have piggybacked and grafted onto the dynamic synergies created by cross fertilization of similar and compatible disciplines. The hybrid economies that have taken shape in Boston and in the San Francisco Bay Area are interesting but they represent slideshots, instantaneous views of many complex dynamics taken place and gelling together. One of these central dimensions is the migration process that brings people together or, alternatively, tears apart communities of scientists. Migrants speed up transformations in places of intense innovations but they can also destabilize institutions, as when an entire department migrates to a different university to enjoy better research conditions.

Migration studies have paid attention to these processes but they have also focused nearly exclusively on cases of migrations framed by national identities: Chinese migrants in the textile industry, Mexican migrants in the low skilled occupations of California or Indian migrants in the IT sector in Silicon Valley. Despite this trend to capture migration processes, there is a lack of data about migration, and more problematic, a lack of understanding of its dynamics. The migration is seen either from the perspective of demographic approach or from the perspective of personal biographies. Both approaches are important. However, they analyze the process on a scale, which

is not always appropriate for understanding of the role of international mobility for specific professional worlds.

In this paper we leverage a database that provides us with information on the location and academic production of Russian computer scientists. This dataset allows us to ask anew a series of questions:

What is the pattern of mobility among Russian computer scientists who work in academia? Are those patterns related to scientific productivity?

Russian computer scientists allow us to focus the larger questions raised by migration studies scholars onto this population in order to understand Russian highly skilled workers mobility.

We explore Russian computer scientists' mobility using bibliographic data as a source of information. We develop a combination of methods which allows us to go beyond a simple description of migration flow or individual trajectory analysis. The source of data is publications indexed in Web of Science, the scientific publications aggregator by Thomson Reuters.

2. DATA

We extracted our data set from the Web of Science using a query combining subject categories and publication years. We formally selected all categories related to computer science:

COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE

COMPUTER SCIENCE HARDWARE ARCHITECTURE

COMPUTER SCIENCE INFORMATION SYSTEMS

COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS

COMPUTER SCIENCE SOFTWARE ENGINEERING

COMPUTER SCIENCE THEORY METHOD

Total number of selected publications is 1499127 for the period of 1985–2013. In this article we analyze only 1990–2012 period.

We parsed and transformed the data to a Sqlite database using CorText platform (http://manager.cortext.net/, [2]). For each record we extracted the countries of affiliation of all authors and their names and surnames. The crucial step was to select most active (3 publications or more) researchers with former USSR nationality. We relate the nationality to individual's name and surname. Name and surname can be an indicator of individual's country of origin. But if a country knows a long history of massive outward international mobility this indicator is not precise: we can find John Smith in UK, USA, Australia, South Africa. In the case of individuals having former Soviet Union origin the chance to be precise is much higher because of absence of large migration flows before 1991. Our name selection algorithm is based on Naïve Bayesian classifier similar to described in [3]. However, some manual check was performed to control the error level, basing on that we can estimate the method as giving about 10% percent of false positive results. After the removal of such false Russian the final sample includes publications of 3656 individuals.

3. METHOD

Bibliographic data is a rich material for understanding how the scientific careers are shaped. However, the career perspective has rarely been applied in exploration of this data, therefore, an appropriate research method has not been established yet. In order to analyze Russian computer scientists' careers, we developed a new approach. We transformed the bibliographic data into a longitudinal form, a sequence of states, and then we explored it using Optimal Matching (OM) method for sequence analysis. After having clustered individual careers, we visualized and described the obtained clusters in order to understand resulting type patterns of individual career of researchers. Furthermore, we calculated and visualized the probability of transition between different states in the total population for each moment of time. That permitted us to seize how the flow of individuals from one type of publication activity to another changes in time.

3.1 Sequence analysis and Optimal Matching

We start the description of our method by presentation of sequence analysis which is the central method for our approach. The term sequence analysis englobes a set of methods which take as a basic unit an ordered sequence of categorical values [1]. The OM method consists in the calculation of degree of pairwise difference (called distance) between such sequences and goes together with few ways to graphically represent sequence data and is often extended by a clustering based on calculated distance data.

There are many algorithms that permit to establish a distance between sequences, most of them, such as Levenshtein's, are familiar to computer scientists. In Social Sciences, sequences are often analyzed with the OM algorithm. It calculates the number of operations, such as insertion, deletion of a state from a sequence and replacement of one state within a sequence by another, which is needed in order to transform one sequence to another. Each operation is characterized by a number called "cost", the distance between two sequences is the minimal possible sum of costs of all operations transforming one sequence to another.

3.2 Transformation of bibliographic data into sequence format

OM can be applied only to sequences. The way we transform bibliographic data into sequence format should take into account our research question. In our case, we question international aspects of researchers' publication activity. Hence, a state in a sequence is supposed to summarize information about international dimensions of individual's publication activity in a given year. For each year of each individual trajectory we list all affiliations indicated in articles published by an individual, including affiliation of co-authors. We summarize affiliation data in three categories: first, there are years where all affiliations are located only in former USSR countries; second, there are years where all affiliations are located only outside those countries; third, there are years where the list of affiliation locations includes both former USSR and foreign countries. Unfortunately, the database does not attribute an affiliation to an author, it can only provide a list of affiliations for all the authors of a given article. Therefore, we cannot seize individual affiliation directly. If in the first and the second case we can deduce information about the affiliation of an individual (former USSR and foreign, respectively), in the third category we mix individuals with foreign affiliations who publish with former USSR located researchers and those who are located in the former USSR but publish with co-authors with foreign affiliations. Despite the fact that the problem concerns 804 individuals, it does not deny us to perform the analysis – mixed

list of affiliation of an article gives us important information about international openness of researcher. Even if we do not know where he works, we know that he has is publishing within a group including researchers affiliated both in former USSR and abroad, or have double affiliation.

The affiliation typology gives us four categories of states. Each state describes particular configuration of publication activity in a given year. First, the most frequent state is the absence of publication in a given year; we label it "NoPub". Secondly, a state where, in a given year, all affiliations listed in publications are foreign. We label it as "ForAf". Thirdly we put all situations where a mixed list of affiliations is observed in a state labeled as "MixAf". Then we observe a situation where affiliations are located only in former USSR we label this state as "RusAf".

Table 1. State distribution in dataset

State label	Number of occurrences
ForAf	12698
MixAf	1683
RusAf	3911
NoPub	65796

The dataset covers the period from 1990 to 2012. The total number of each state in the dataset is presented in Table 1.

3.3 Cost setting and clustering

Having presented sequence data, we are proceeding to optimal matching analysis. We set costs of operations following the basic settings implementation of the algorithm in the TraMineR package for R [3]. We used substitution costs based on transition rates from one state to another. According to Katherine Stovel, "The assignment of transformation costs haunts all optimal matching analyses" [7:394]. However, we tried different ways of cost setting and discovered that distance matrix and resulting clustering do not vary a lot depending on the chosen method, which often happens in case of well-structured data[1]. Having established a distance matrix between individual sequences, we perform descendant hierarchical clustering using Ward's algorithm.

4. RESULTS

4.1 States and transitions

The information on state distribution in studied population trajectories describes a large context of individual activity. Table 1 shows that the most frequent publication activity configuration for computer scientists from the former USSR is, except the absence of publication, to have a foreign affiliation, we have three times less publications with only former USSR affiliation, the mixed affiliation is about seven times less frequent then the foreign one. Therefore, the point of departure for our analysis is the fact that the majority of former USSR computer scientists works abroad and has no visible scientific ties to the homeland.

In the Figure 1 we summarized the annual distribution of states for the studied population. We omit the most frequent state, the absence of publication, in order to make other state frequencies more visible. The main trend is the growth of publication activity which concerns all the states. Independently of affiliation location, the number of publications roughly increased from 1990 till 2009 with a temporary drop around 2006, and then we observe a permanent drop from 2009 to 2012. The fastest growth concerns publications with foreign affiliation, we supposing to observe a discipline expansion at international level from 1999 to 2007. The former USSR affiliated publications number seems to grow less dramatically. The fall in late years is due to objective fall in publication activity and by our database artifact: it was established in late 2012 and excludes some new publications.

The number of states with a foreign affiliation solely is growing slower during early nineties and is even deposed by number of states with former USSR affiliation in 1994. It could be explained either by the fact that some of the journals published in the former USSR were included in the WOS database and/or by a more active publishing of research papers in English. In any case, these two explanations could have in common a supposition that this grow is based on research done before the collapse of the Soviet Union. Then we observe a decrease in 1995-1996, the moment of the lowest level of R&D employment due to fall in public support to science [5]. Afterwards the growth of homeland affiliated publications is never as fast as of foreign affiliated ones – the effect of crisis of Russian science seems to be quite strong.

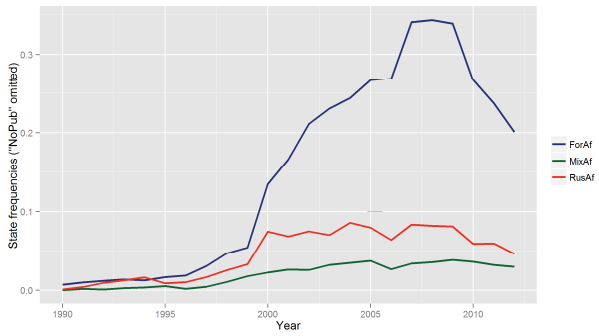


Figure 1. State distribution per year

The second step of analysis is to explore the temporal dynamic of transition processes. We analyze how the probability that an individual switches from on type of publication affiliation configuration to another is changing in time. We calculate transition matrices which indicate the probability of transition from one state to another within a year. The studied period lasts 22 years and there are four distinct states. Therefore, the data about temporal evolution of transition rates is presented in 22 4x4 matrices. Such information is not easy to analyze. In order to do so, first of all, we present in Table 2 a 4x4 matrix representing mean transition rates for all the period. The rows and columns represent respectively departure and arrival states, transitions happening rarer than two times per year are excluded from calculation.

Researchers with foreign affiliations have a quite stable profile – they keep publishing in the same configuration, the probability of the change to another pattern (except the absence of publication) is of 3%. Former USSR affiliation is a less stable case: its change probability is of 8 %. Such a state is also characterized by the highest probability of the absence of publication activity the following year; therefore, former USSR affiliated researchers are less active in general. The state "mixed

[1] For an exhaustive review of cost-setting issues in social sciences see [6].

affiliations" is characterized by multidirectional exchange flows. Also the mixed affiliation pattern gives the highest probability to have publication activity the following year. The mixed affiliation state has an intermediate role: the transition rates to other states are very high. Partially it could be explained by the fact that publishing with co-authors having foreign affiliation may be correlated with further researcher's international mobility. But also it is caused by the fact, that the state, as described in section 3.2, bears some ambiguity on individual's location. Still it is a good indicator of international openness of researcher's activity.

Table 2. Transition rates

	[->NoPub]	[->RusAf]	[->MixAf]	[->ForAf]
[->NoPub]	83.2%	4.3%	1.3%	11.2%
[->RusAf]	65.7%	26.3%	5.7%	2.3%
[->MixAf]	49.0%	11.2%	19.6%	15.5%
[->ForAf]	51.4%	0.9%	2.1%	45.5%

The transition flow is not permanent. In Annex 1. we represent in a graphic form the annual deviation of transition rates. The graph is not standard and should be read as follows. We have four graphs representing transition rates for four "departure" states. Blue and red strips connect the "departure" state in a given year with a "destination" state the following year. These strips visualize the deviation from mean transition rate; the end of the line position indicates "destination" state, it value is shown by tick marks on the left. The color of the strip represents the sign of deviation (red is positive, blue is negative) its thickness represents it relative value. For example, we are looking for data about transition rates from states with mixed affiliation. We will find "MixAf" label on the right of the second graph from the bottom. Than we choose the "departure" year, referring to the axis in the bottom of the graph, for example, 2000. The thin red strip starting at 2000 x-coordinate and "MixAf" y-coordinate and ending at the point with 2001 x-coordinate and "RusAf" y-coordinate shows that in 2000-2001 period the probability to switch from mixed affiliation to former USSR affiliation was somewhat greater than average probability to do such a switch within the overall period. However, two years later the same switch happened even more frequently, the red line starting at 2002 x-coordinate and "MixAf" y-coordinate and ending at 2003 x-coordinate and "RusAf" y-coordinate is consequently thicker.

We can observe that the switches between affiliation configurations tends to grow in the period of active publication. Years where transitions from "No Publication" state to other, "active" states is above average are characterized by transitions between other states. This observation brings us to a conclusion that for the overall population the publication activity correlates with the change of country of affiliation of researchers or/and of their co-authors. But it does not necessarily mean that one determines the other: growing probability to publish and growing probability to change publication configuration can concern different groups in the observed population. This hypothetical tie between switches in publication configuration and scientific productivity will be analyzed from another angle in the next section. We will see a connection between the longitudinal publication patterns and the citation level, which seems to be a more important indicator of productivity than the number of publications.

4.2 Clusters

We preformed clustering with OM algorithms and established 23 clusters. We started by two clusters and stopped when subsequent clustering started producing almost identical clusters. These clusters are presented in the Annexes 2, 3 and 4. Clusters are ordered according to the "scientific success" indicator, the number is indicated above each cluster. To measure the "scientific success", we took a proportion of articles which belong to top ten percent of most cited within this year within a cluster[2]. We also show the number of individuals within a cluster, an average number of articles per author and top four of affiliation countries[3]. We attribute a special color to every state: for example, foreign affiliation state has blue color, mixed affiliation has blue, etc. Annex 2 represents clusters by just plotting all individual sequences within a cluster. The distribution of states for each year (a so-called chronogram) is illustrated in Annex 3. In Annex 4, for each cluster we plot sequences which represent at least 75% of sequences within a cluster[4].

The type of sequences grouped within a cluster is related to clusters "scientific success" level in a quite interesting way. First, we see that clusters where in most states individuals are affiliated to a former USSR country (##20-23) are the least successful. Surely, in part this result could be explained by the fact that individuals actively publish, but only in Russian, but still we can suppose that this cluster represents former USSR computer scientists who have international scientific recognition definitely below the average.

Secondly, clusters where in most states individuals have a foreign affiliation are characterized by at least an average or even, in most cases, above the average success level. Even taking in account the language biases, over-scoring English publications citation levels, we still can suppose that foreign affiliated computer scientists produce scientific results that give them important recognition.

But the most interesting results concern clusters, where we observe a big number of states with mixed affiliations and switches from one publication configuration to other. Such clusters have average or, in most of cases, above the averages success level. The cluster #1 is the most successful: it is characterized by the dominance of mixed affiliation, as can be seen in Annex 3, and by trajectories, including, first, a switch of former USSR based researcher publishing only with co-authors affiliated in the same region to another publishing configuration, especially to publishing articles with mixed list of affiliations. Second, we see there individuals who switch from "only foreign affiliations" configuration to mixed one. Clusters #15 and #9 are

[2] Such a number seems to be a more appropriate measure of "scientific success" than an average citation per article within a cluster, because we neutralize the effect that old articles cumulate more citation with the course of time.

[3] If there is less than four countries that means that other countries are presented in less than 15% of cluster affiliations, it here is a + sign it meant that there is more than four countries presented on more than 15% of cluster affiliations.

[4] That means that at least 75% of a cluster sequences have a distance to representative sequences which is less than 10% of maximal theoretical distance between sequences within a dataset. Refer to [4] for more details about representative sequences.

analogous to cluster #1 in terms of patterns structure, but different in terms of position in time. We can deduce that described patterns correlate with a slightly above average scientific success in the case of early careers in computer science and show outstanding success only in the case of newcomers

Giving attention to cluster #5 we cans see quite exceptional situation: many publication patterns last since 1990 and are characterized by dominance of former USSR affiliations and by presence of French affiliation. However, the performance is above average. We cannot offer an exhaustive explanation for this observation, but we suppose that we observe here a strong former USSR-based research community, its success being related to achievements of Soviet science in a certain field.

5. DISCUSSION

A great part of analyzed trajectories is quite stable. The characteristics of publication activity tend not to change often. In the case of former USSR-located researchers, this stability correlates with a low citation level, foreign-affiliated researchers tend to be well-cited. At the same time, researchers switching affiliation are also well cited. Moreover, after 2005, scientific careers engaging changes in the country of affiliation of an individual and/or of his co-authors are leading to a greater scientific success than less turbulent careers. Researchers with complex internationally open careers seem to be quite productive and successful (but not the best) when they start to work before or during discipline expansion (1999-2007). For those who started later, such type of career became a clue to leadership in terms of citation levels.

6. CONCLUSION

Computer scientists who originate from the former USSR and publish in Web of Science referred journals have, in most of the cases, foreign affiliation. Those few of them who have affiliation only in the former USSR and have homologous co-authors have poor citation level in most cases. So, international dimensions of scientific activity are an important factor of scientific success. But there is no best way to internationalize the publication activity. Even if the migration with stable publication pattern when researcher is working and publishing abroad seems to be related to a high citation level the difference with, firstly, foreign-affiliated scientists publishing with ex-compatriots, and secondly, with scientists who change publication patterns during their career is not so important. Moreover, for the overall population, growing publication activity is correlated with a growing rate of change of publication configuration and, after 2005, the most cited articles are made by scientists with turbulent careers.

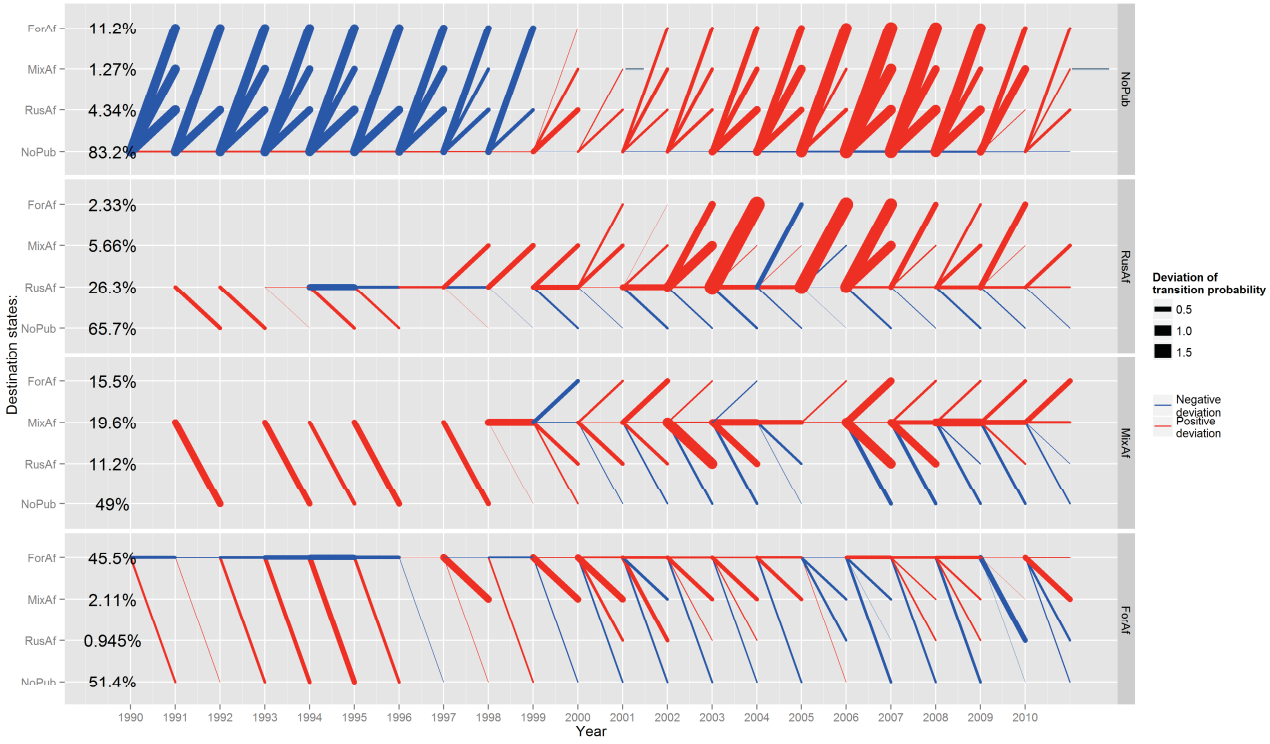
The internationalization of former USSR computer scientists can be seen as a process when researchers located in an area characterized by relatively poor research funding and relatively small number of active colleagues try to accede, by collaborating and migrating, countries where the domain is strong and well founded. But this view is simplistic: scientists and institutions located in the former USSR seem to bring fruitful collaboration for researchers with foreign affiliation. We can observe complex patterns of relations between diaspora scientists and their homeland research universe and complex patterns of internationalization of foreign USSR based researchers. Those complex patterns are related to higher scientific success. Described dynamics concerns only some part of researchers; nevertheless, they show that the unidirectional “brain drain” vision of former USSR scientist’s migration should be nuanced.

On the methodological level, our claim is that sequence analysis with optimal matching could be a good method to explore scientific career through bibliographic data.

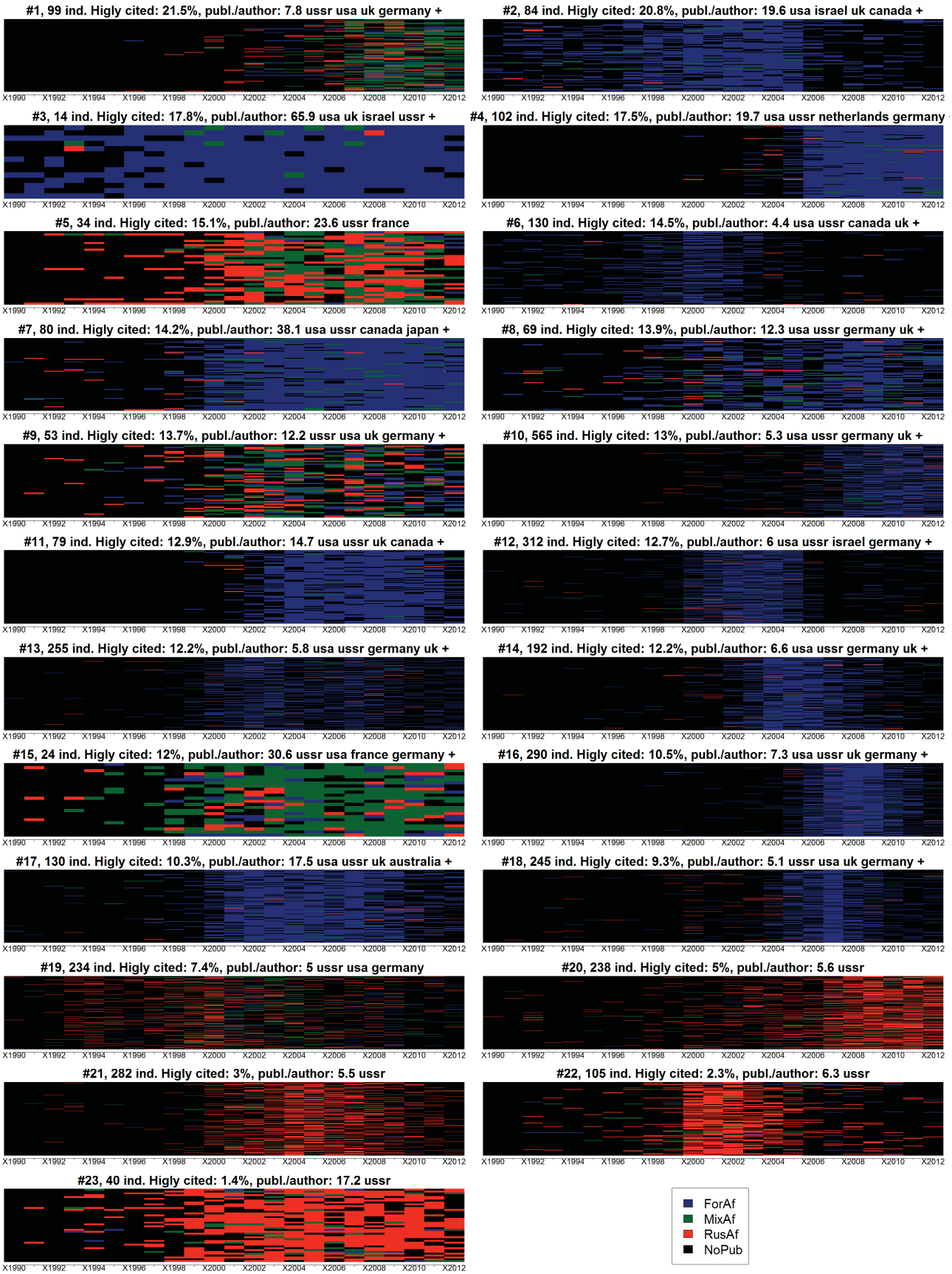
7. REFERENCES

1. Abbott, A. and Tsay, A. Sequence Analysis and Optimal Matching Methods in Sociology. *Sociological Methods & Research 29*, 1 (2000), 3 –33.
2. Andrei Mogoutov, Cambrosio, A., Keating, P., and Mustar, P. Biomedical innovation at the laboratory, clinical and commercial interface: A new method for mapping research projects, publications and patents in the field of microarrays. *Journal of Informetrics 2*, 4 (2008), 341–353.
3. Flores, J.J.G., Zweigenbaum, P., Yue, Z., and Turner, W. Tracking Researcher Mobility on the Web Using Snippet Semantic Analysis. In H. Isahara and K. Kanzaki, eds., *Advances in Natural Language Processing*. Springer Berlin Heidelberg, 2012, 180–191.
4. Gabadinho, A., Ritschard, G., Studer, M., and Müller, N. Extracting and Rendering Representative Sequences. In A. Fred, J. Dietz, K. Liu and J. Filipe, eds., *Knowledge Discovery, Knowledge Engineering and Knowledge Management*. 2011, 94–106.
5. Graham, L.R. and Dezhina, I. *Science in the new Russia : crisis, aid, reform.* Indiana University Press, Bloomington, 2008.
6. Lesnard, L. Using optimal matching analysis in sociology: Cost setting and sociology of time. (2014).
7. Stovel, K., Savage, M., and Bearman, P. Ascription into Achievement: Models of Career Systems at Lloyds Bank, 1890-1970. *American Journal of Sociology 102*, 2 (1996), 358–399.

Annex 1. Annual deviation from mean transition rates for each departure state

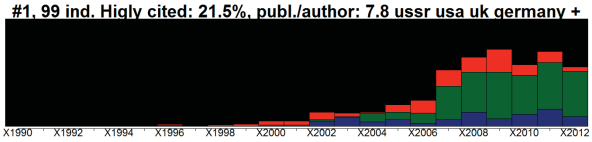


Annex 2. Clusters of individual sequences

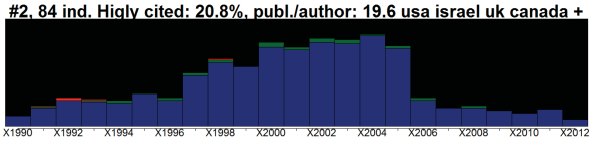


Annex 3. State distributions for clusters

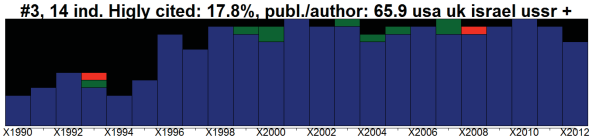
#1, 99 ind. Higly cited: 21.5%, publ./author: 7.8 ussr usa uk germany +



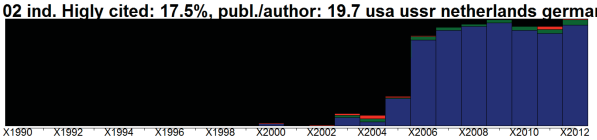
#2, 84 ind. Higly cited: 20.8%, publ./author: 19.6 usa israel uk canada +



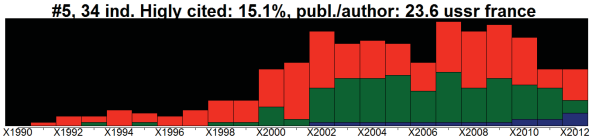
#3, 14 ind. Higly cited: 17.8%, publ./author: 65.9 usa uk israel ussr +



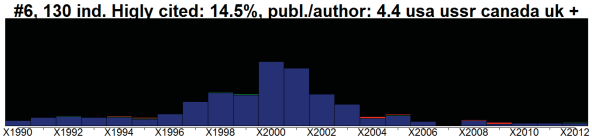
#4, 102 ind. Higly cited: 17.5%, publ./author: 19.7 usa ussr netherlands germany



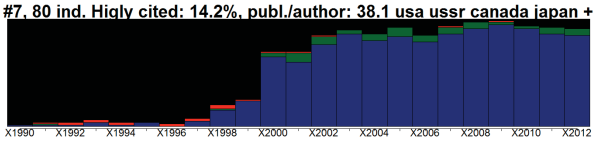
#5, 34 ind. Higly cited: 15.1%, publ./author: 23.6 ussr france



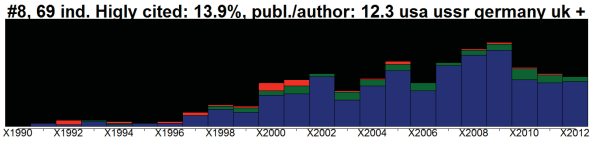
#6, 130 ind. Higly cited: 14.5%, publ./author: 4.4 usa ussr canada uk +



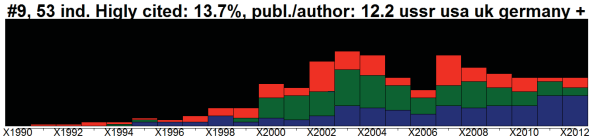
#7, 80 ind. Higly cited: 14.2%, publ./author: 38.1 usa ussr canada japan +



#8, 69 ind. Higly cited: 13.9%, publ./author: 12.3 usa ussr germany uk +

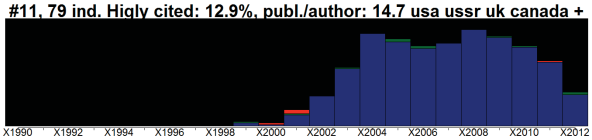


#9, 53 ind. Higly cited: 13.7%, publ./author: 12.2 ussr usa uk germany +

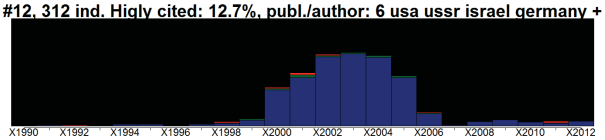


#10, 565 ind. Higly cited: 13%, publ./author: 5.3 usa ussr germany uk +

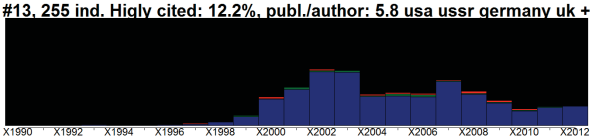
#11, 79 ind. Higly cited: 12.9%, publ./author: 14.7 usa ussr uk canada +



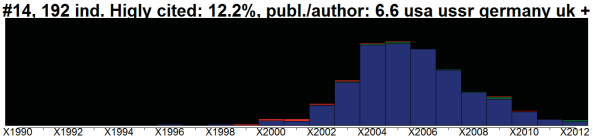
#12, 312 ind. Higly cited: 12.7%, publ./author: 6 usa ussr israel germany +



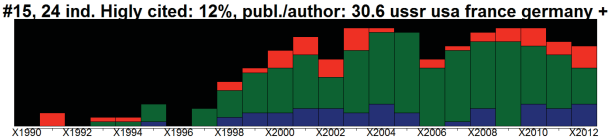
#13, 255 ind. Higly cited: 12.2%, publ./author: 5.8 usa ussr germany uk +



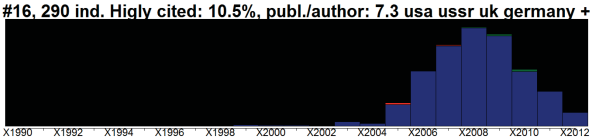
#14, 192 ind. Higly cited: 12.2%, publ./author: 6.6 usa ussr germany uk +



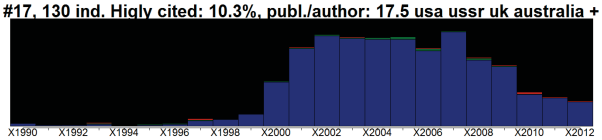
#15, 24 ind. Higly cited: 12%, publ./author: 30.6 ussr usa france germany +



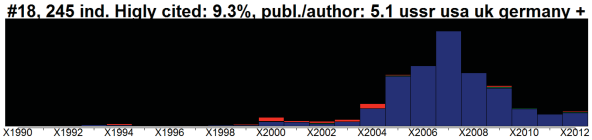
#16, 290 ind. Higly cited: 10.5%, publ./author: 7.3 usa ussr uk germany +



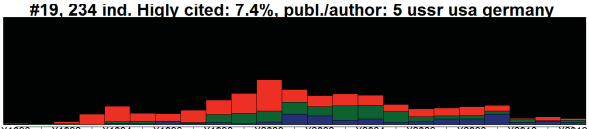
#17, 130 ind. Higly cited: 10.3%, publ./author: 17.5 usa ussr uk australia +



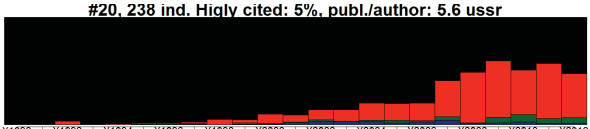
#18, 245 ind. Higly cited: 9.3%, publ./author: 5.1 ussr usa uk germany +



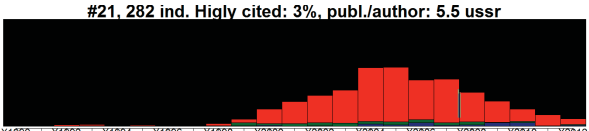
#19, 234 ind. Higly cited: 7.4%, publ./author: 5 ussr usa germany



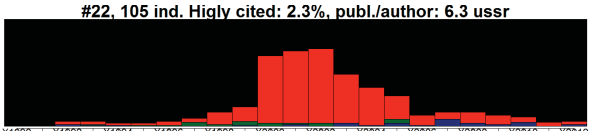
#20, 238 ind. Higly cited: 5%, publ./author: 5.6 ussr



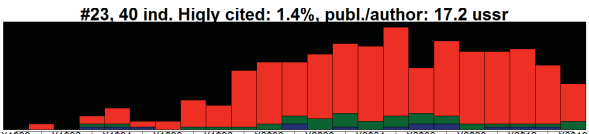
#21, 282 ind. Higly cited: 3%, publ./author: 5.5 ussr

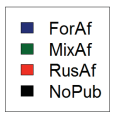


#22, 105 ind. Higly cited: 2.3%, publ./author: 6.3 ussr



#23, 40 ind. Higly cited: 1.4%, publ./author: 17.2 ussr





Annex 4. Representative sequences for clusters