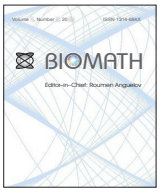
Original article Biomath 1 (2012), 1209031, 1–7

Biomath Forum

BIOMATH

http://www.biomathforum.org/biomath/index.php/biomath/



Exploring Family Relations between International Patent Applications

Peter Hingley
European Patent Office
Erhardtstrasse 27, Munich, Germany
Email: phingley@epo.org

Received: 12 July 2012, accepted: 3 September 2012, published: 10 October 2012

***Abstract*—In the international system for granting patents for inventions, first patent filings can be followed by subsequent filings at other patent offices within one year. Each such group of related filings constitutes a patent family. Tests are developed as to whether the observed number of first filings that leads to subsequent filings (*r*) is in agreement with a random process of assignment of the hits from the subsequent filings. An exact expression for the random distribution can be used for small sized data sets. Its behaviour and also the behaviour of an asymptotic Poisson approximation as well as a censored binomial distribution for *r* are assessed. The approach is stimulated by the Fisher-Wright model in population genetics and possible parallel applications to other biological processes are sought, such as transformations of stem cells and cancer.**

***Keywords*-censored binomial; genetics; patents; random assignment**

I. Introduction

Usually an inventor starts a quest for intellectual property protection by making a first patent filing at the local national patent office. Then, within one year, subsequent filings quoting the priority of that first filing can be made at any patent office. These are termed subsequent filings. Unlike most national patent offices, the applications that are received at the European Patent Office (EPO) are mostly subsequent filings, due to its supranational character as an umbrella Office for the European Patent convention contracting states (EPC), which also have their own national offices to which applications can be made [1].

To aid the statistical description of the flows of such (provisional) patent rights, the concept of *patent families* is useful. These are explained in II. The PRI database is a patent families file that is extracted from a worldwide patent database at EPO called DOCDB, that itself contains data on patent publications from all the main offices around the world [2]. A subset of the documents in DOCDB represents published patent filings that can be identified as representing either first filings or subsequent filings, depending on whether or not they contain priority references to earlier first filings. In PRI the data are re-ordered and compacted so that each record is indexed by a priority reference. Information is also given on the activities of subsequent filings that quote that priority, such as the major geographical blocs in which subsequent filings took place.

Studying the international spread of patent filings combines concepts from several streams. Mainly since the 1960s there have been studies of patent economics and statistics, starting with examples of the patenting process as motivators for econometric models, but later on centring more on elements of the system itself that has become an important economic driver [3]. In parallel the subjects bibliometrics and scientometrics have been developed. Network theory can also be relevant [4] because the relationship of subsequent filings to first filings is not one-to-one, even when considering just a single first filing office / subsequent filing office pair. That is, one first filing can be quoted as a priority in more than one subsequent filing, and a subsequent filing can quote several first filings as priorities.

Citation: P. Hingley, Exploring Family Relations between International Patent Applications, Biomath 1 (2012), 1209031, http://dx.doi.org/10.11145/j.biomath.2012.09.031

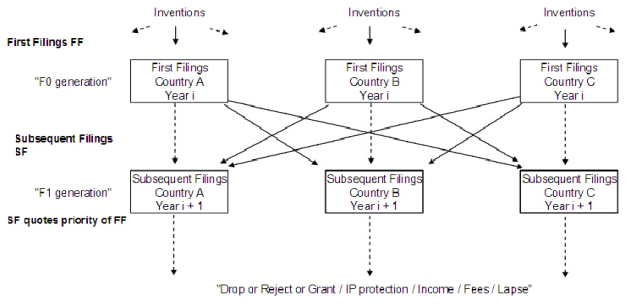


Fig. 1. Structure of patenting in terms of a generation of first filings, followed by subsequent filings up to one year later.

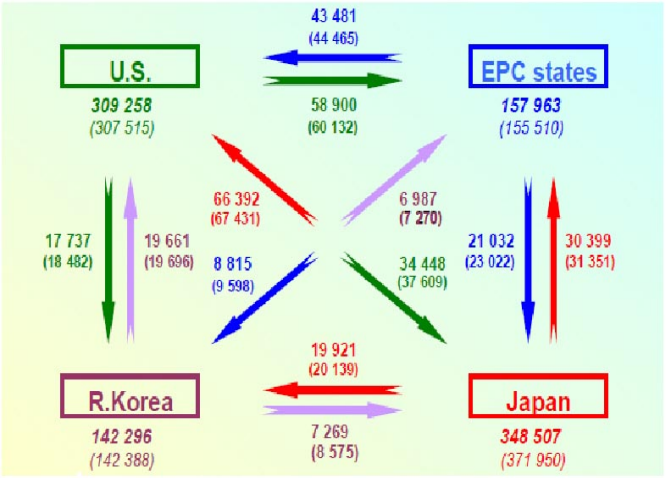


Fig. 2. Single priority patent families arising from first filings in 2006 (2005 in brackets), indicating first filings and flows, which are the counts of first filings referenced as priorities in subsequent filings in other blocs. From [7].

Here some probabilistic aspects will be considered, concentrating on subsequent filing activities at the EPO. The concepts to be explored are inspired by tests for random mating in population genetics. Patent family populations do not directly exhibit so many directly analogous characteristics to biological populations, but nevertheless it is interesting to study the parallels and differences. It will be suggested that the methods might have some as yet unrecognised ability to model cellular processes in biology, or at least be able to give fresh insight into experiments and models that could be tried out.

II. The Structure of Patenting and Patent Families

First patent filings (FFs) lead on to subsequent filings (SFs) in other countries up to one year later. Imagine two generations as set up in Fig. 1, rather like the Fisher-Wright model in population genetics [5]. The FFs are the *F0* generation and the SFs are the *F1* generation. The members of F0 do not all reproduce, but some do to give one or more offspring in F1. Each member of F1 however must have at least one parent in F0.

The parallels with biology do not go much further than this in any strict sense, because after F1 the patents are examined and granted if considered worthy, then maintained against the payment of appropriate fees for up to 20 years before they lapse. This means that no further reproduction of this cohort can normally take place. The generational pair of populations F0 and F1 can be said to be renewed over and over again every year. (This could be paralleled perhaps in population genetics by a model for pets.) Beyond the limited set-up considered here however, the population of inventors and firms that make patent applications persists over time as well as with dynamic entries and departures each year [6]. The act of filing for patents can be considered as a possible survival tactic in a competitive world. There are no genes or DNA in patent families, although there are technical classification systems to describe the areas covered by a patent that play some kind of analogous role.

There are various types of patent families according to different definitions. For *single priority families*, which will be used here, each family constitutes one FF together with all the SFs that lead from it. Thus each FF from which a priority filing emanates can initiate one family only. But the SFs in F1, that are the offspring of the FFs, can belong to more than one family.

More extensive definitions of patent families are possible, that include for instance *composite patent families*, where each family consists of a complete interconnected network of FFs and SFs. This has the advantage of making every family unique, because no patent publication can belong to more than one family. However this may not be such an important consideration. In a study involving the whole population of recorded publications with earliest priorities in the period from 1991 to 1999, it was found that the nine most common family structures relate to a single priority and make up more than 77 per cent of patent families [8]. Also, 29 per cent of families consisted of only a pair of one FF with one SF.

Single priority families can be used to describe patent

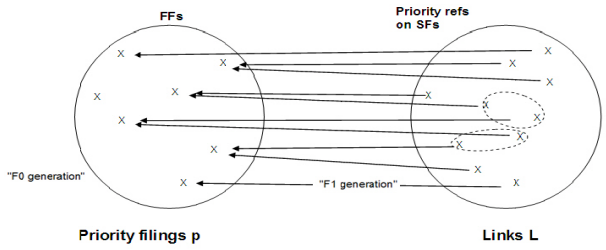


Fig. 3. Random hits model.

filings flows between countries (Fig. 2), although the subtleties of multiple assignments between FFs and SFs should also be taken into account to give a complete description. In order to do this, the concept of *links* is important. That is, each priority reference from a SF forms a link to the FF that is referenced. In a set of p FFs and y SFs emanating from those FFs, involving in each case filings at one or more patent offices, let there be L links. Say that the average number of SFs that are linked to a FF is ϕ, while the average number of FFs that are linked to a SF is θ. These averages relate to L as follows [2].

$$L = p.\phi = y.\theta$$

III. Random Hits Model for Priorities

Fig. 3 shows that the setup of p priority filings in the F0 generation with L links to y SFs in the F1 generation can be represented by a surjective directed graph. For the idealised model to be considered here, the occasional groupings of several members of L into sets that represent SFs with several priority references will be ignored.

Let r be the number of members of p that are hit by the L links. If p and L are considered to be fixed, r can be modelled by a random process of hits on p by L.

A. Exact Distribution

Feller [9] developed the following formula for the discrete probability distribution of r, under the hypothesis of a process of independent random hits.

$$Pr(r) = \frac{1}{p^L}\binom{p}{p-r}\sum_{\nu=0}^{r}(-1)^{\nu}\binom{r}{\nu}(r-\nu)^L$$

This is valid for any values of L and p that are positive integers. No explicit expression is given for the moments of this distribution. Feller's examples concentrate on the case $L > p$, such as where r is the number of days in a year when there is at least one birthday in a village of 2000 people ($L = 2000$, $p = 365$). In our case here $L < p$, because only a proportion of FFs lead to SFs.

For the following calculations, routines were written in R. $Pr(r)$ is easily computable only when r is small. Fig. 4 shows $Pr(r)$ for the case $p = 30$ and $L = 20$. Direct evaluation gives a mean of 14.77 and a standard deviation (square root of variance) of 1.49. The distribution was checked by constructing the r values obtained in one million simulated sets of data, where each set was formed by sampling randomly with replacement the first p integers L times. The resulting histogram is indistinguishable visually from Fig. 4, with a mean of 14.76 and a standard deviation of 1.50. This shows good agreement with the exact distribution.

B. Poisson approximation

Feller [9] argues for a Poisson approximation for $Pr(r)$ as p and $L \longrightarrow \infty$. Say t is the number of members of p that do not lead to subsequent filings. He asserts that, if $\lambda = pe^{-\frac{L}{p}}$ remains bounded,

$$Pr(t) \longrightarrow e^{-\lambda}.\frac{\lambda^t}{t!} \qquad | \qquad [0 < t < \infty]$$

The support of this distribution is not bounded above. In fact p is finite and we are interested in $Pr(r = p - t)$. This can be approximated by a transform of the Poisson distribution, bounded above at p but unbounded below 0.

$$Pr(r) \approx e^{-\lambda}.\frac{\lambda^{(p-r)}}{(p-r)!} \qquad | \qquad [-\infty < r \leq p]$$

Fig. 5 shows $Pr(r)$ for the case $p = 30$ and $L = 20$, and can be compared to the exact distribution in Fig. 4. Direct evaluation gives a mean of 14.61 and the quantity $p - \lambda$ is 14.60, showing good agreement with the mean according to Feller's argument, and not too far from the exact distribution mean of 14.76. However the standard deviation is 3.92, which is more than twice as high as 1.49 for the exact distribution. The shape is also different, and is essentially censored at the upper limit of 20, where $r = L$.

C. Censored Binomial Approximation

Since the Poisson approximation does not work well at this sample size, other approximations can be tried. A censored binomial distribution is in some way equivalent

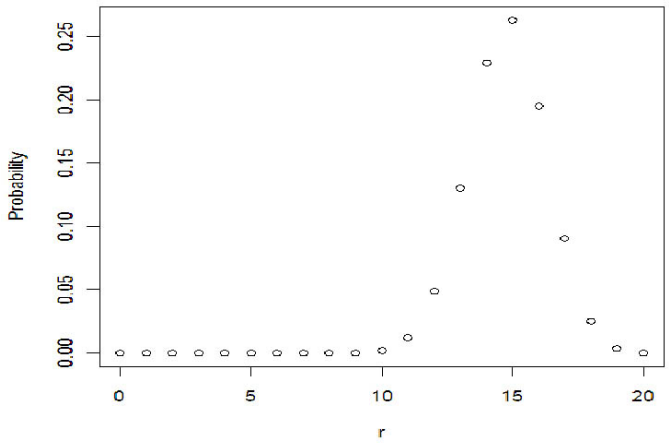


Fig. 4. Exact distribution with $p = 30$ and $L = 20$.

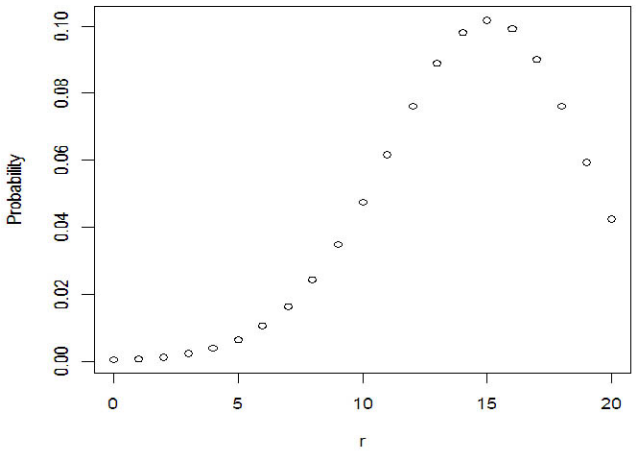


Fig. 5. Poisson approximation with $p = 30$ and $L = 20$.

to Feller's exact distribution, except that the hit probabilities are all considered independent and equivalent, and dependency due to the conditional probability chain is ignored.

Say that s is the number of hits from L to a member of p. This can be represented approximately as $Binomial(s, L, \frac{1}{p})$, meaning the binomial probability of s successful outcomes when there are L independent trials, each with probability $\frac{1}{p}$ of success.

$$Pr(s) \approx \frac{L!}{s!(L-s)!}(\tfrac{1}{p})^s \left(1 - (\tfrac{1}{p})\right)^{(L-s)} = Binomial(s, L, \tfrac{1}{p})$$

Under an assumption of independence, the probability that a particular member of p is hit is then as follows.

$$1 - Binomial(0, L, \tfrac{1}{p}) = 1 - (1 - \tfrac{1}{p})^L$$

r is the number of distinct members of p that are hit. If $L \geq p$, as in the birthdays example in IIIA,

$$Pr(r) \approx \frac{Binomial(r,p,1-(1-\frac{1}{p})^L)}{[1-Binom(0,p,1-(1-\frac{1}{p})^L)]}$$

This is a censored binomial that removes the zero class, because at least one member of p must be hit ($Pr(0) = 0$).

But in the patent families case, where $L < p$, the response range is restricted to r in $(1, ..., L)$, so there is also censorship to remove all classes between and including $L + 1$ and p.

$$Pr(r) \approx \frac{Binomial(r,p,1-(1-\frac{1}{p})^L)}{[1-Binom(0,p,1-(1-\frac{1}{p})^L) - \sum_{j=L+1}^{p} Binomial(j,p,1-(1-\frac{1}{p})^L]}$$

Fig. 6 shows $Pr(r)$ for the case $p = 30$ and $L = 20$, to compare with Fig.s 4 and 5. The R routine in this case calculates the probabilities using the normal approximation to the binomial distribution. It was checked that this makes minimal difference to usage of the exact binomial expressions, even at this small population size.

Direct evaluation gives a mean of 14.63, again fairly close to the exact distribution mean of 14.76. This time the standard deviation is 2.64, which is closer than the Poisson approximation to the 1.49 for the exact distribution. The shape is however still quite different to the exact distribution, although not as far away as the Poisson was.

IV. Random Hits Model for Patent Families Data

The distributions in III can be scaled up to give tests of random hits to patent families with SFs at EPO. In the following examples, FFs at EPO were ignored because they were already hit in a sense at the time of first filing. It should also be recognised that FFs and SFs at EPO do not represent all the patenting activity in Europe, because of the alternative possibility to file at the national patent offices in each EPC contracting state. Note also that the analysis will be monospecific, in that it is only the flows to EPO that are considered, and not the spread of flows to all offices, as was considered in [8].

In order to scale up to the case of an annual data set of first filings (F0 generation) and the subsequent filings that quote them as priorities (F1 generation), consider p =1 052 420 worldwide first filings in the year 2002 and L =135 439 references to these priorities that were made in SFs to EPO, mainly in the year 2003. The subset

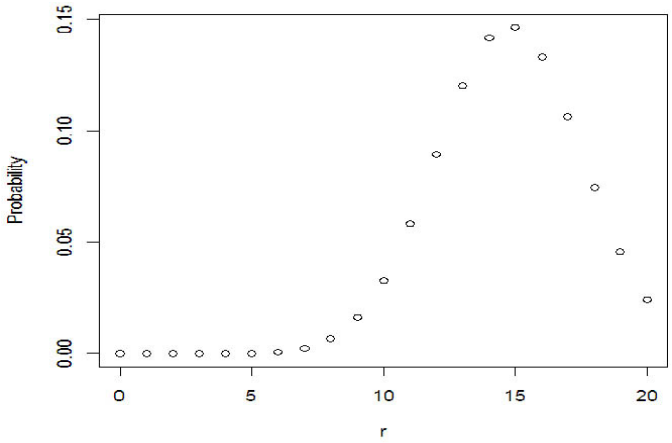


Fig. 6. Censored binomial approximation with $p = 30$ and $L = 20$.

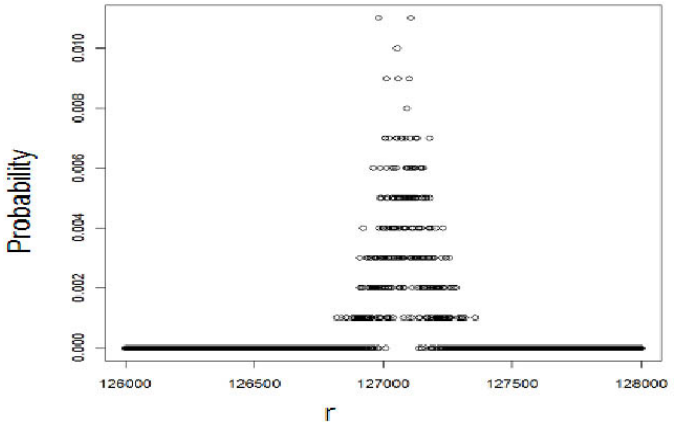


Fig. 7. Simulated distribution of r, based on 1 000 simulated data sets with $p = 1\ 052\ 420$ and $L = 135\ 439$.

of p that were referenced was of size r =120 701. How near was this to a random process of hits of L on p?

Feller's exact distribution is not calculated in this case because it is no longer straightforward to do so with large numbers (the practical upper limit for the R routine is about $p = 60$). Simulations are also more time consuming, but it is possible to make enough of these to get a good idea of the shape of $Pr(r)$. Fig. 7 shows a first estimate of the distribution that was made with 1000 simulated data sets. What is being emulated here is presumably a unimodal discrete distribution like Fig. 4, but due to a lack of binning we see a discretised approximation ($Pr(r) = 0.001$ equivalent to one simulated outcome, $Pr(r) = 0.002$ equivalent to two simulated outcomes, etc.). The mean is estimated as 127 079 and the standard deviation is 84. While these parameters are obviously not determined with great accuracy, due to the small number of simulations carried out, the shape of the distribution indicates that the observed value $r =$ 120 701 is significantly lower than its expectation under the random hits model.

The simulation results in Fig. 7 lie in a very tight range around their mean, compared to the support. Under a normal approximation, 95 per cent of the simulated r values are expected to be between 126 911 and 127 247.

The distributions $Pr(r)$ according to Feller's Poisson formula and the censored binomial approximation are shown in Fig.s 8a and 8b respectively. They are both centered close to the mean of the simulations, and $p - \lambda$ from the Poisson approximation is 127 086, which is also close to the mean of the simulations. But the spreads of both distributions are again too large, with standard deviation according to the Poisson distribution at 962 and for the censored binomial formula at 334. However it can be seen in this figure that the observed value of r is still significantly too low to be entirely random, even for the Poisson formula.

So it seems that the number r of worldwide priorities in 2002 that were hit by EPO SFs was lower than expected under a random hits model. The test was also carried out on priorities after separation into the main geographical blocs of origin (EPC, Japan, US, Others) and over five priority years (2002 to 2006 inclusive). See Figs. 9a to 9d. The expected values under the random hits model are represented in these diagrams by values of $p - \lambda$ (triangles).

The results are fairly consistent over the years that were studied. There are less hits than expected for priority references to US first filings, but more hits than expected for priority references to EPC, which is the European home area for EPO operations. This suggests that there is only a subset of US FFs that somehow qualify for filings as SFs later on at EPO, which is reasonable for a large country with some of its own specific internal markets that are not relevant abroad. For Europe, the contrary result means that priorities are better sampled than expected and rarely lead to multiple EPO SFs. The results show r conforming more or less to its expectation under random hits for Others origin and Japan origin priorities (in the case of Japan at least for the years 2004 to 2006). The values of $\frac{L}{p}$ (average number of links to priorities overall) differ between blocs (EPC 35 per cent, Japan 8 per cent, US 20 per cent, Others 2 per cent, for priorities in 2002). It is interesting that in Japan and Others this was far less than in the other two blocs. Perhaps the fact that the probability of a hit was lower has led to a better fit of the Poisson

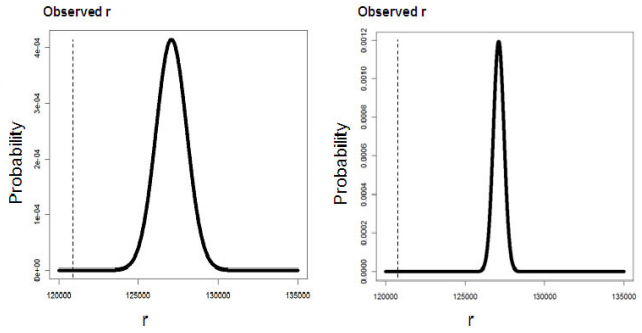


Fig. 8. a) Poisson approximation; b) Censored binomial approximation; with $p = 1\ 052\ 420$ and $L = 135\ 439$.

approximation or to the random hits model in general.

V. Possible Applications in Biology

In II it was suggested that temporal F0/F1 generations with annual replacement represents only a special case in population genetics. The schema lacks the attractive equilibrium properties that are interesting when making theoretical predictions about population dynamics and evolution over many generations.

But there may be possibilities for applications of such models to special cases in biology. Consider for example the transformation of an undifferentiated bank of stem cells into tissues and/or organs, such as the development of clones of immune cell against specific antigens [10]. How many of the stem cells will differentiate and into what tissue, if there are indeed different possible stem cell fates in development? In III and IV an unexpected tightness of the distribution $Pr(r)$ was found around its mean. This suggests that the number of stem cells committing to become certain organs under a random differentiation model may be almost constant, even if random. Aberrations in the process could perhaps occur in cancer.

It may be useful to study competition between tissues as sinks for stem cells. In the patent world this is analogous to studying the fates of first filings in terms of priority references from subsequent filings in several other offices. For example such counts appear in [7] in terms of trilateral (EPO, Japan, US) and Four Office (EPC, Japan, Korea, US) family subsets.

Another extension to the present model that can be beneficial to consider in both biological and patent regimes is the case where there are several conversions of the original entity via a sequence of transforming hits taking place in a temporal series. In the patent world there is the sequence of transformations of the priority forming first filing into a subsequent filing, followed by the possible grant of the patent and its eventual expiry, not to mention the collection of a cumulative set of fees at the patent office in lieu of these various steps. In biology there are sequences of cellular development that lead down limited paths of development under certain restrictions, such as colon crypt cell growth. This is a special case to which population genetics theory can be adapted, and brings us back towards schemes such as in Fig. 1 [11].

VI. Conclusions

The development started with a description of the family relations of groups of patents in terms of population genetic parameters, and then continued by developing specific tests of random assignment of subsequent filings to priorities via distributions of hits. It turned out that the distributions are so tight that the outcomes almost appear to be fixed, even though the underlying process is random. There was a brief consideration of how biological models for certain special phenomena may be able to make use of the method.

The exact formula in IIIA gives the best representation of the effects of random hits, but it is not easily calculable when constructing a null distribution for large data sets. Feller's argument for a limiting Poisson distribution does not apply well for the case that $L < p$, because its variance is too large. However the quantity $p - \lambda$ is a good approximation for the mean. A censored binomial distribution is also well centered and has a lower variance than the Poisson, but is still too disperse. A closer approximation to the exact distribution should be established, that can work with larger numbers and stays as close to the original formula as possible.

For patent families that involve subsequent filings at EPO, the observed number of priorities is less than that predicted by the random hits model. This is mainly due to less hits than expected from applicants in US, although there are more hits than expected from Europe. To model these situations more explicilty, it may be beneficial to develop a weighted version of the exact formula, where combinations with fewer hits have higher weights (US case) or lower weights (EPC case) than combinations having a greater number of hits.

Apart from the possible extensions that were mentioned in V, it will also be interesting to develop more intricate models of the international patenting system. This could include an extension of the model presented here to test independence of hits to a common set of first filings when subsequent filings are made to several

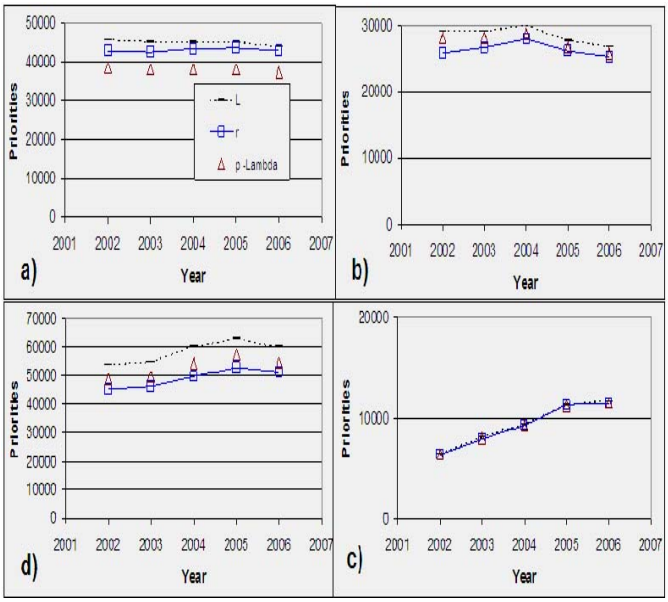


Fig. 9. The results for apparent randomness of r depend on blocs of origin of the patent families. Reading clockwise from top left: a) EPC, b) Japan, c) Others, d) US. L: dotted line; r: boxes; $p - \lambda$: triangles.

other patent offices. Models could also be developed to consider the effects of having several priority references to different first filings from some of the subsequent filings.

ACKNOWLEDGMENT

The author would like to thank colleagues for supplying and maintaining the PRI data file, and Marc Nicolas for useful discussions on the methods.

REFERENCES

[1] P. Hingley, and M. Nicolas, (eds), *Forecasting Innovations, methods for predicting numbers of patent filings*, Springe, 2006.
[2] P. Hingley, *Patent families defined as priority forming filings and their descendents*, http://forums.epo.org/students-2-students/topic720.html (2010).
[3] Z. Griliches, *Patent statistics as economic indicators: a survey*, Journal of Economic Literature 28, 1661–1707 (1990).
[4] J.C. Vivar, and D. Banks, *Models for networks: a cross-disciplinary science*, WIREs Comp. Stat. 4, 13–27 (2012). http://dx.doi.org/10.1002/wics.184
[5] J. Ewens, *Mathematical population genetics*, Vol. 1, 2nd edition, Springer, 2004. http://dx.doi.org/10.1007/978-0-387-21822-9
[6] P. Hingley, and S. Bas, *Numbers and sizes of applicants at the European Patent Office*, World Patent Information 31, 285–298 (2009).
[7] European Patent Office, Japan Patent Office, Korean Intellectual Property Office, United States Patent and Trademark Office. *Four Office Statistics Report, 2010 edition*, http://www.trilateral.net/statistics/tsr/fosr2010.html (2011).
[8] C. Martinez, *Insight into different types of patent families*, http://www.oecd.org/dataoecd/21/32/44604939.pdf (2010).
[9] W. Feller, *An introduction to probability theory and its applications*, Vol. 1, Wiley, 1968.
[10] D. Wodarz, *Killer cell dynamics*, Springer, 2007. http://dx.doi.org/10.1007/978-0-387-68733-9
[11] M.A. Nowak, *Evolutionary dynamics*, Belknap Harvard, 2006.