Time to First Citation Estimation in the Presence of Additional Information

Tina Nane

g.f.nane@cwts.leidenuniv.nl
Centre for Science and Technology Studies (CWTS), Leiden University, P.O. Box 905, 2300 AX, Leiden (The Netherlands)

Abstract
We are interested in modelling the time to first citation, that is how long does it take for a publication to be cited for the first time after it has been published in a journal. We argue that both cited and uncited publications should contribute to the distribution of the time to first citation. Moreover, our objective is to model the time to first citation nonparametrically, hence under no parametric assumption. Due to the similarities with the observed data in survival analysis, we employ the techniques based on censored data and describe the distribution of the time to first citation in terms of the hazard rate, that is the instantaneous rate of being firstly cited. We find that publications receive their first citation at increasing rates in the first 24 months after their publication date and at decreasing rates afterwards. Moreover, we observe that the hazard rate and hence the time to first citation is influenced by the document type, number of authors and collaboration type and field. We also investigate the difference in the time to first citations for publications grouped by their collaborative status or the assigned field.

Conference Topic
Citation and co-citation analysis

Introduction
The first citation a publication receives is an important event in the bibliometric data, as it is not only a simple citation count, but also marks a change in the status of the publication, i.e. from being uncited the publication becomes cited. Certainly, observing the first citation of a publication depends on the considered time frame. Regardless the period of analysis, certain publications will never receive their first citation, in other words we will not observe the first citation received by some publications for any finite time period we consider.

Another important aspect concerns the time it takes for a publication to receive its first citation. For some publications it takes a small amount of time, such as 1-2 months, while for others it can even take more than 10 years. Due to overlong reviewing and publication procedures, some publications might even have negative times to first citation, meaning that the publication has been cited before it has been published.

The event that a publication received its first citation, as well as the time to the first citation received considerable attention over the years, starting with Schubert and Glänzel (1986), Glänzel (1992), Rousseau (1994), Glänzel and Schoepflin (1995). Since 2000, Egghe (2000), Egghe and Rao (2001), Burrell (2001), and Glänzel et al. (2012) continued to model the first citation data. Additionally, we acknowledge the work of van Dalen and Hekens (2005) and Bornmann and Daniel (2010), that is specifically close to the present research and will be referred to later on. Most of the previous work relies on the parametric modelling of the time to first citation distribution, such as the double exponential model (Rousseau, 1994), mixtures of non-homogeneous Poisson process

(Burrell, 2001), etc. The modelling in the existing literature focuses only on publications in certain journals or fields and the uncited publications do not always contribute to the time to first citation distribution, yet they emerge as a consequence of the model (Burell, 2001). Additionally, in Egghe (2000), the proportion of the uncited documents emerges from the model.

It should be stressed however that the time to first citation distribution derived from a set of publications that contains both uncited and cited documents does not coincide with the time to first citation distribution of the publications that receive a citation. From a probabilistic perspective, the first distribution is the sub-distribution of the latter. Furthermore, not accounting for the uncited publications can lead to biases in the estimation of the distribution of the time to first citation.

Our present study aims to continue and extend the research on the time to first citation analysis. We consider all the publications, regardless the document type and field, that appeared in Web of Science (WoS) in 2000 and their first citations received until the end of 2013. The time to first citation is registered in months. Additional data is recorded for each publication, such as document type, the number of authors, institutions and countries, and information on collaboration.

We propose an approach that aims to model the time to first citation distribution by accounting for all observations (both uncited and cited publications). Our approach assumes that the event of interest is the first citation, which is time dependent and we are interesting in modelling the time to this event of interest, namely the time to first citation. The time to event analysis has been employed in many fields. In sociology, it is known as event history analysis, in economy as duration analysis and in engineering is called reliability theory. Nevertheless, it is best known in biostatistics, where most research has been performed and where it is called survival analysis.

Consequently, the terminology employed in survival analysis is ubiquitous. In biostatistics, a frequent event of interest is death and the time to the event is then expectedly called survival time. Different functionals of the distribution of the time to the event of interest are successively termed survival function, hazard or cumulative hazard function. We will employ this unfortunate terminology in the analysis of the time to first citation.

A typical feature of the data in survival analysis is that not all events of interest are observed within the period of analysis. These observations are referred to as censored observations. The uncited publications are therefore regarded as censored observations. The uncited publications are in fact right censored observations, since their first citation is conditioned to take place after the period of analysis ended, i.e. at the right of the period of analysis. This approach circumvents the issue of not having a time to first citation for the uncited publications.

In survival analysis, the distribution of the time to event data is usually characterized by its survival function, as well as its hazard rate. The hazard rate provides information on the evolution in time of the event rate, in our case first citation rate. An attractive feature of the hazard rate compared to the density function, for example, is that the hazard rate accounts for the aging effect, while the density does not. Based on our data, we provide the time to first citation distribution and investigate its behaviour via the hazard rate.

Another important aspect in survival analysis is how additional information on observations, referred to as covariates or explanatory variables influence the time to the

event of interest. The Cox model (Cox, 1972) is probably the most popular method to model the influence of covariates on the time to the event of interest. In this study, we aim to infer on the effect of different characteristics of publications on the time to first citation. In other words, is the document type, number of authors, collaboration type or the field of a publication influencing the time it takes for that publication to receive the first citation? To our best knowledge, the influence of the explanatory variables document type, collaboration or field have not been accounted so far in the time to first citation analysis.

These methods in survival analysis have been previously used to model the time to first citation distribution by van Dalen and Henkens (2005) and Bornmann and Daniel (2010). Both studies restrict themselves to publications in a specific area of research, i.e. demography and chemistry. van Dalen and Henkens (2005) propose to model the hazard rate of the time to first citation distribution under the parametric assumption of a Gompertz distribution, which, in turn, lead to hazard rate which are decreasing over time. This restriction is unintuitive and in particular, it does not fit the data of the present study. Bornmann and Daniel (2010) are very brief in explaining the methods and, more importantly, the results of the analysis are not consistent in presenting their results, as they first refer to the differences in the survival curves and later on to the differences in the hazard rate. It is not very clear, for example, if the publication characteristics have an effect on the hazard rate.

Time to first citation distribution

We consider all the publications in Web of Science (WoS) that appeared in 2000 and their first citations up until 2013. That accounts for 1,202,371 publications, from which 62.62% received their first citation until the end of 2013. The first citation of publication A is defined as the publication date (month) of a publication B that cites firstly publication A, that is the minimum publication date of all publications that cite publication A. Needless to say that since the study is restricted to WoS, we refer to the first citation covered by WoS. Moreover, we exclude self-citations, hence we condition on publication B having no common authors with publication A.

The time to first citation of publication A is the time period (in months) between the publication date of publication A and the publication date of a publication B that cites firstly publication A. The time to first citation can sometimes be negative, but this is mostly an artefact due to the slow reviewing or publication process in different journals, etc. We exclude such observation from our study.

We chose the publication date to be registered in months given the availability of the data, but also for a better insight in the first citation process. Moreover, this avoids the issue of highly discrete data. Nonetheless, it is noteworthy that the publication date in months is not available for all data. For these cases, the first month of the year (January) or the middle one (July) is usually reported.

The histogram of the time to first citation for the publications in WoS that appeared in 2000 and received their first citation within the period 2000-2013 is presented below.

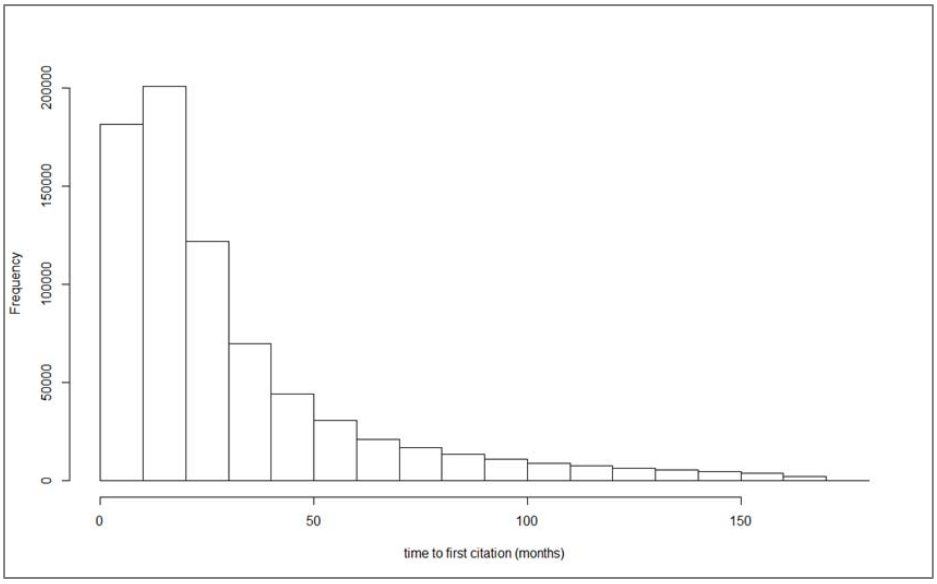

Figure 1. Histogram of the time to first citation for publications in 2010.

Most of the publications received their first citation shortly after publication. As expected, the proportion of publications that receive citations decreases over time. There are however publications that receive their first citation 13 years after their publication.

The histogram provides information on the time to first citation distribution of publications that received at least a citation until 2013. As mentioned beforehand, there is however no information on the publications that have not received any citation, apart from the percentage of the uncited publications.

Censored observations

It would be desirable though that the uncited publications also contribute to the distribution of the time to first citation, as they influence the probability of being firstly cited. Within this framework, the uncited publications did not experience the event of interest (first citation) by the duration of the study. What it is known is that their first citation occurs after the analysis ended.

In survival analysis, these observations are referred to as right censored observation. The publications that received their first citation within the period of analysis are called uncensored observations. Modelling time to event data requires that observations, both censored and uncensored have an observed time of interest, denoted as the follow-up time. For the uncensored observations, the follow-up time is the time to their first citations. For the censored observations, the follow-up time is the time period (in months) between their publication date and the end of analysis, that is December 2013, and it is referred to as the censored time.

For example, the censored time of a publication that appeared in January 2000 is 168 moths, whereas the censored time of a publication from June 2000 is 163 months. It needs to be distinguished between a publication with its time to first citation 163 months, for example it appeared in January 2000 and was firstly cited in December 2013, and a publication with its censored time 163 months. For this, we use an indicator Δ that is 1 if the publication has been cited and 0 if the publication remains uncited for the period of analysis.

The hazard rate

We are now interested in modelling the first citation rate on small units of time and its evolution in time. For this we will make use of the hazard rate, a functional of the time to first citation distribution. The hazard rate is referred to as the force of mortality in sociology, or the failure rate, in reliability. All these terms adhere to the pessimistic tone consistently used in survival analysis.

The hazard rate quantifies the rate at which first citations occur per unit of time relative to the proportion of publications that have not been yet cited. For a continuous random variable X, the hazard function is defined as

$$\lambda(t) = \lim_{\Delta t \searrow 0} \frac{P(t \leq X < t+\Delta t | X \geq t)}{\Delta t}.$$

In our case X denotes the time to first citation. We assume that the underlying time to first citation is continuous, while the observed data is discretized by measurement.

In order to compute the hazard rate at a given time point t, one needs to calculate the conditional probability in the numerator. In the present study, this is the probability of being firstly cited in the time interval [t,t+Δt), given that the publication has not been cited before time t. The conditioning ensures that at each time point t, only the publications that have not been cited up until time t are considered, therefore also the publications that are not cited throughout the entire period of analysis, i.e. the censored observations. Dividing this conditional probability by Δt, that is the width of the interval [t,t+Δt), we obtain the rate of the first citation occurrence per unit of time. By taking the limit Δt↘0 gives the instantaneous rate of occurrence of first citation. Note that, by definition, the hazard rate is not a (conditional) probability, or a density.

The hazard rate is a functional of the time to first citation distribution and can be derived for any parametric distribution and also estimated for a nonparametric distribution. The most straightforward example is the exponential distribution, for which the hazard rate is a constant function.

The hazard rate for the publications in the study is depicted in Figure 2 below.

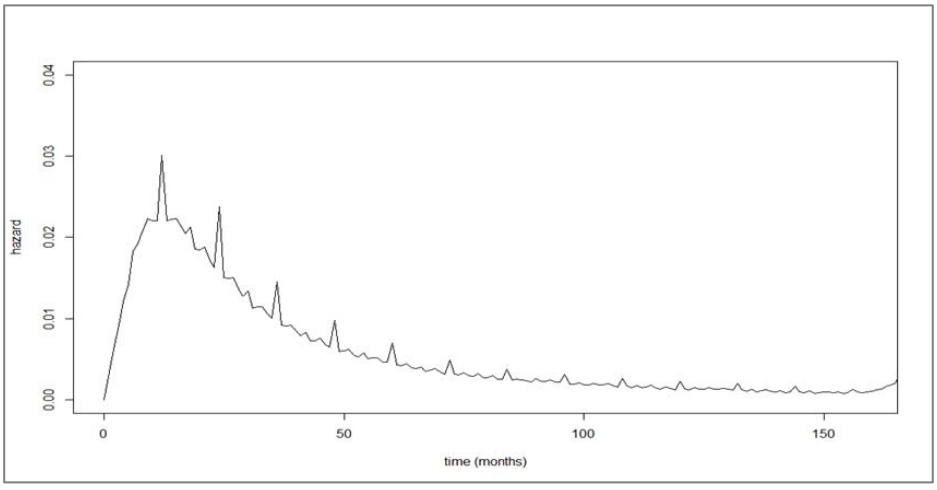

Figure 2. Hazard rate of publications in 2010.

First of all, we notice some spikes in the hazard function, which occur at the beginning and in the middle of each year in the citation window. This is due to the fact that certain journals publish once or twice a year. Moreover, when the publication date of certain

journal issues is unknown, the publication date is typically assigned to the beginning or middle of the year.

It seems that, per unit of time, publications receive their first citation at an increasing instantaneous rate up until a given time, that we refer to as the first citation peak, and despite the spikes, at decreasing instantaneous rates after the first citation peak. This shape suggests an unimodal hazard rate.

The first citation peak is for this dataset 24 months. In terms of conditional probabilities, the results can be interpreted as follows. Given that publications have not been cited before, on small unit intervals, they get cited for the first time with higher probability in the first 2 years after publications and with lower probability afterwards. The conditional probability decreases with time, but flattens after a while. That is, the decrease of the hazard is rather steep until 50 months and flattens afterwards. It can be inferred that first citation instantaneous rate is low and does not change significantly for documents that have not been cited for 4-5 years after publication.

Additional information – covariates

We are now interested in what can possibly influence the time to first citation and its hazard rate. This additional information is recorded as explanatory variables that are typically referred to as covariates in survival analysis, or as control variables in econometrics.

We consider the following covariates: document type, number of authors, collaboration type and field. By field we refer to the 250 subject categories to which journals are assigned in WoS. Surely, other covariates might be included, such as number of institutions or countries, number of pages, journal impact, etc.

Assume that covariates do not change over time, that they have a fixed value at the publication date. There can be however, covariates that change over time (time dependent covariates), such as journal impact, authors' visibility or performance.

The Cox model

The most famous model that incorporates the information on certain covariates in survival analysis is the Cox model (Cox, 1972). Regardless the fact that the model is more than 40 years old, it has been widely used and numerous versions, for particular issues with the data, have been proposed and investigated ever since.

The Cox model specifies the hazard rate at time t of a publication with a given covariate vector z as

$$\lambda(t|z) = \lambda_0(t)\exp(\beta' z),$$

where λ_0 is the underlying baseline hazard and β' is the transpose of the vector of underlying regression coefficients. Notice that if we take all covariates to be zero, we obtain the baseline hazard.

Within the Cox model, the hazard has two components. The first one, the baseline hazard, is the nonparametric part and it indicates how the hazard varies in time. The second term specifies parametrically, via an exponential function, the dependence on the covariates. It is then obvious why the Cox model is considered a semi-parametric model. Moreover, it is worth mentioning that the baseline hazard can be left unspecified when one want to estimate the regression coefficients and this flexibility has been particularly attractive for researchers.

Ever since the model was proposed, there was a great interest in estimating the regression coefficients β, that reflect how changes in the covariates produce a change in the hazard rate. The estimates were obtained via a partial likelihood method that avoided the bothersome issue of estimating the baseline hazard λ_0.
We have fitted the Cox model with the following covariates

- Document type
- Collaboration type
- Number of authors.

We will focus on estimating the (baseline) hazard and not on the regression coefficient estimation. We need to stress that conditioning on the covariates to be at a baseline value, i.e. z=0, is not the same thing as not accounting for covariates. This can be determined from the equation specifying the Cox model, but also from the figure below.

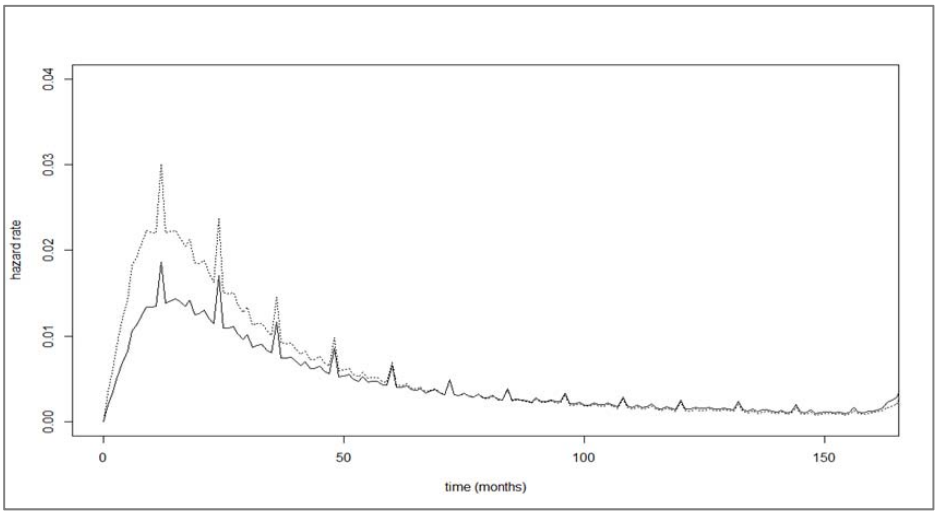

Figure 3. Hazard rate in the presence of no covariates (dotted) and baseline hazard (solid line).

Apparently, accounting for covariates shifts the hazard down in the first 60 months after the publication date and has no effect afterwards. The baseline hazard follows the same trend as the hazard rate in the presence of no covariates that is increasing until 24 months after the publication date and decreasing afterwards. Therefore, we can conclude that the covariates have a scale effect rather than a shape effect on the hazard. Furthermore, it seems that there is a proportional effect of the covariates on the baseline hazard, at least in the first 50 months. This represent a visualization of the goodness of fit of the Cox model and additionally, several tests suggest that the model fits the data well.
We want to investigate now whether certain characteristics of the publication, such as the collaborative status or the field have an impact on the instantaneous first citation rates.

Collaboration

It is commonly thought that publications that have resulted from an international collaboration are more visible to the academic community and hence receive more citations than national collaborative publications or publications that do not result from any inter institutional collaboration. It would be interesting to see if the collaboration type also influences how fast a publication receives its first citation.

As mentioned beforehand, we have fitted a Cox model with document type, collaboration type and number of authors as covariates. All the covariates have a (statistical) significant influence on the time to first citation.

To show the difference in the hazard rates among the different types of collaboration, we compute the hazard rate for publications with international, national and no collaborations. All the other covariates are set to their baseline level. Figure 4 depicts these differences.

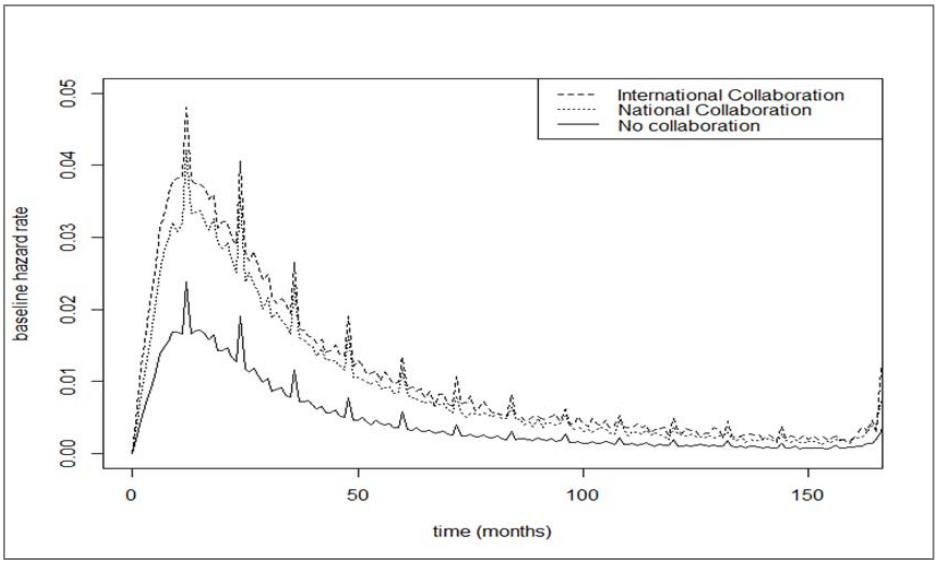

Figure 4. Baseline hazard rates in terms of collaboration type: international collaboration (dashed), national collaboration (dotted) and no collaboration (solid line).

It seems that there is a significant scale difference in the instantaneous first citation rate among publications that represent international and international collaborations and those that do not result from such collaborations. There are however small differences between baseline hazard of the international and national collaborative publications. Nonetheless, the publications that resulted from an international collaboration register higher instantaneous first citation rates than publications that represent national collaborations and these publications have, in turn, higher instantaneous first citation rates than publications whose authors are affiliated to a single institution. Similar to the overall (baseline) hazard rates, there are less and less differences in the hazard rates of different collaboration types 100 months after publication.

Contrary to the popular belief however, it seems that, apart from a scaling factor, publications receive their first citation at similar rates irrespective their collaboration type. The maximum hazard function is attained by publications of all collaboration types at the same time point, which is 24 months after the publication date. This is not different from the overall baseline hazard.

To condition further on specific values of the other covariates, we have considered the document type ‘Article’ and assume the publications has 3 authors, which is close to the overall average of the entire dataset, that is 3.31.

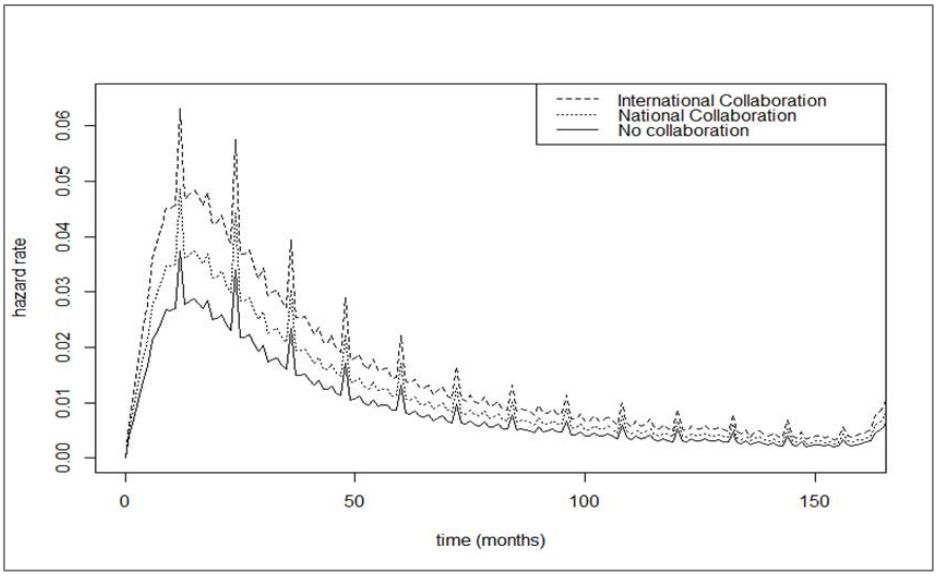

Figure 5. Hazard rates for articles with mean number of authors. International collaboration (dashed), national collaboration (dotted) and no collaboration (solid line).

Figure 5 depicts the hazard rates of articles that result from different collaborations and are written by three researches. We notice that the differences in the hazard rates have decreased. Despite similar behaviour over time, international collaborations still achieve the highest hazard rates over time, followed by national collaborations and articles produced by the same institution (no collaboration).

Field

We are also interested to see whether the field assigned to a certain publication affects the rate of being firstly cited. Nonetheless, more than half of the journals in WoS are assigned to at least two fields and some journals are assigned to six fields. This means that the field covariate cannot be uniquely defined for each publication. This difficulty cannot be overcome by using the WoS subject category assignment and hence the field cannot be included as a covariate in the Cox model. A solution is to adopt the publication-level classification system proposed by Waltman and van Eck (2012). Within this approach each publication is assigned to an unique cluster. Employing the publication-level classification system is deferred to future research.

In order to still assess the influence of the field on the time to first citation distribution, we have limited the data of all publication from 2000 to three fields: Biochemistry & Molecular Biology, Economics and Mathematics. We have now a number of 80,745 publications that have been published in 2000 and are assigned to the three fields.

We have fitted the Cox model with the following covariates

- Document type
- Collaboration type
- Number of authors
- Field

All four covariates have a (statistical) significant effect on the hazard rate. We are interested in the baseline hazard rates for the data grouped by the field. The differences

between the three baseline hazards can be observed in Figure 6. Once again, the other covariates have been set to zero.

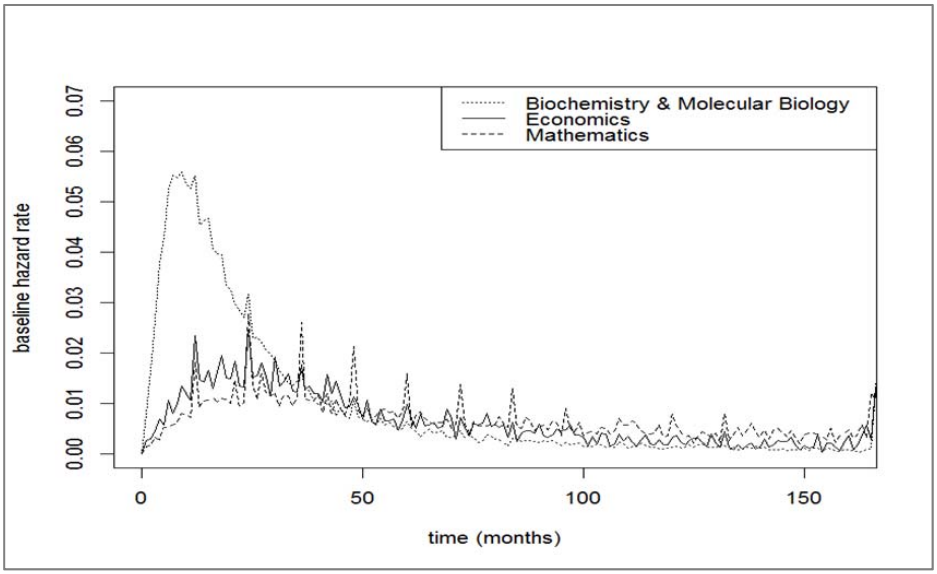

Figure 6. Baseline hazard rates in terms of field: Biochemistry and Molecular Biology (dotted), Mathematics (dashed) and Economics (solid line).

The three baseline hazard rates differ in both shape and scale. Firstly, it seems that the publications that appeared in 2000 in Biochemistry and Molecular Biology achieve their maximum first citation rate earlier than publications in Economics or Mathematics. The peak in Biochemistry and Molecular Biology is registered at 12 months, whereas the publications in Economics and Mathematics have a baseline hazard rate peak around 24 months.

We observe that there are large changes over time in the baseline hazard rate of publication in Biochemistry and Molecular Biology. Moreover, during the first part of the citation window, publications in Biochemistry and Molecular Biology have an instantaneous first citation rate three times as higher than the instantaneous first citation rates in Economics and Mathematics. The publications in Economics and Mathematics exhibit similar hazard rate behaviour.

It is noteworthy and interesting that after 60 months, the order of the three baseline hazard rates completely reverse, that is publications in Mathematics have higher baseline hazard rates than publications in Economics, that have higher baseline hazard rates than the publications in Biochemistry & Molecular Biology.

Discussion and conclusions

The first citation is probably the most important citation a publication receives. It can determine entirely the number or speed of further citations. Besides a simple citation count, it also changes the status of a publication, from being uncited to being cited. In some fields, being cited is even sufficient to become frequently cited.

The time to first citation also contributes to the number or speed of further citations. Apart from the famous sleeping beauties (van Raan, 2004), it is obvious that the more it takes for a publication to receive its first citation, the lower the probability of receiving further citations.

Time to first citation is the first step in modelling how publications accumulate citations in general over time. It is still unknown whether the time to first citation differs significantly from the time to second citation, etc.

We aimed to model the time to first citation and used a set of publications that appeared in 2000 and are included in the WoS database. Probably the most important aspect of our approach is that we employed nonparametric or semi-parametric methods of estimation. In other words, we let the data speak for itself. This ensures a greater flexibility and avoids the bothersome issue that a given model fits a particular data well, say publications that appear in a certain year and within a specific field, but fails to fit another particular data appropriately. While this is not a problem specific only to the first citation analysis, for an example on this matter in the first citation analysis, see Rousseau (1994). Another important drawback of the parametric approach is that certain employed parametric models cannot incorporate specific shapes of the first citation data. Van Dalen and Hekens (2005) for example make use of a Gompertz hazard model that cannot incorporate an unimodal hazard, as we obtained in the present study.

Apart from the nonparametric choice of estimation, we have also incorporated the uncited publications in the distribution of the time to first citation by using methods developed in survival analysis. We stress the fact that the information on uncited publications should be accounted for in modelling the time to first citation distribution, otherwise the results of the estimation can be seriously biased, especially given the high percentage of uncited publications.

We have investigated the time to first citation distribution through its hazard rate, the instantaneous rate of being firstly cited. We observe that the hazard rate increase over the first 24 months and decreases afterwards. This is somehow expected, that publications receive their first citations at higher rates until a maximum and afterwards at lower and lower rates. What is surprising is the relative short period of time over which the hazard rate is increasing. It means that the probability of a publications being cited for the first time is increasing over the first 24 months, and decrease afterwards.

Furthermore, it is of high interest to investigate whether certain characteristics of publications influence their time to first citation. We included the document type, number of authors, collaboration type and the field. We have found that all these explanatory variables (covariates) influence the time to first citation and investigated the differences between the hazard rates of publications grouped by collaboration type. The hazard rates of the three collaboration types differ in scale and not in shape and attain the maximum at the same time point. Hence, it seems that publications receive their first citations at an increasing rate up to the same time point, namely 24 months regardless their collaboration type.

A different dataset has been chosen to investigate the influence of the field on the time to first citation. It seems that, for the three selected fields, the hazard rate of the publications differ not only in scale but also in shape. The publications in Biochemistry and Molecular Biology register higher rates than publications from Economics and Mathematics, but also they have increasing first citation rates over a shorter period of time than the publications from the other two fields. The order of the three hazard rates reverse after 60 months.

As mentioned in the previous section, the problem of the overlapping fields in WoS needs to be addressed in future research and this can be overcome by considering the

publication-level classification system proposed by Waltman and Van Eck (2012). Numerous investigations are further required and desired. For example it would be very interesting to investigate whether the time to first citation distribution, and in particular the hazard rate including self citations differs from the time to first citation excluding self citations. Other covariates can be included in the analysis, such as the impact of the journal, the performance or visibility of authors, etc. Of course, it is very interesting to see whether the shape of the hazard rate changes over the time of publication, not only through the citation window. The author expects that the hazard would have the same unimodal shape, but the maximum point would be attained at different time points that is the first citation peak would be time dependent.

In terms of estimation, it is highly desirable to account for the monotonicity of the (baseline) hazard that is to provide estimates of the baseline hazard rate under the assumption of monotonicity. This is in line with the research of Lopuhaä and Nane (2013), but needs some refinement to incorporate the estimation of a unimodal baseline hazard.

References

Bornmann, L. & Daniel, H.-D. (2010). Citation speed as a measure to predict the attention an article receives: An investigation of the validity of editorial decisions at *Angewandte Chemie International Edition*. *Journal of Informetrics, 4*, 83-88.

Burrel, Q.L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics,* 52, 3-12.

Cox, D.R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society*. *Series B (Methodological)*, *45*, 187-220.

van Dalen, H.P. & Henkens, K. (2005). Signals in science – On the importance of signaling in gaining attention in science. *Scientometrics*. *64*, 209-233.

Egghe, L. (2000). A heuristic study of the first-citation distribution. *Scientometrics,* 48, 343-359.

Egghe, L & Rao, I.K.R. (2001). Theory of first-citation distributions and applications. *Mathematical and Computer Modelling*, *34*, 81-90.

Glänzel, W. (1992). On some stopping times of citation processes. From theory to indicators. *Information Processing & Management, 28*, 53-60.

Glänzel, W., Rousseau, R. & Zhang, L. (2012). A visual representation of relative first-citation times. *Journal of the American Society for Information Science and Technology*, *63*, 1420-1425.

Glänzel, W. & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature, *Journal of Information Science*, *21*, 37-53.

Lopuhaä, H.P. & Nane, G.F. (2013). Shape constrained nonparametric estimators of the baseline distribution in Cox proportional hazards model. *Scandinavian Journal of Statistics, 40*, 619-646.

van Raan, A.F.J. (2004). Sleeping beauties in science (short communication). *Scientometrics*

Rousseau, R. (1994). Double exponential models for first-citation processes. *Scientometrics, 30*, 213-227.

Schubert, A. & Glänzel, W. (1986), Mean response time – a new indicator of journal citation speed with application to physics journals. *Czechoslovak Journal of Physics (B), 36*, 121-125.

Waltman, L. & van Eck, N.J. (2012). A new methodology for constructing a publication-level classification system in science. *Journal of the American Society for Information Science and Technology*, *63*, 2378-2392.