DOI: 10.18267/j.pep.353

NONPARAMETRIC APPROACH TO PATENT CITATIONS

Petr Mariel, Susan Orbe*

Abstract:
The present article reexamines some of the issues regarding the benchmarking of patents using the NBER data base on U.S. patents by generalizing a parametric citation model and by estimating it using Generalized Additive Models (GAM) methodology. The main conclusion is that the estimated effects differ considerably from sector to sector, and the differences can be estimated nonparametrically but not by the parametric dummy variable approach.

Keywords: USPTO, patent benchmarking, GAM

JEL Classification: O3, C14.

1. Introduction

Patent data are widely recognized as an important source for analyses on innovation, R&D and technical changes *(e.g*. Basberg, 1987; Griliches, 1990). The book by Jaffe,Trajtenberg (2002*b*) is without doubt one of the major elements encouraging widespread use of patent data in economic research, as it includes a huge data base on U.S. patent data and lists the main papers which analyse these data. These data, also available on the National Bureau of Economic Research (NBER) website (www.nber.org) represent the culmination of long research and major effort by many researchers and institutions, as described in Hall, Jaffe & Trajtenberg (2002). This patent data base includes, along with other information, almost three million U.S. utility patents granted between January 1963 and December 1999 together with all citations of these patents made between 1975 and 1999. The NBER is working on an update and extension of this data, bringing existing data up to date through December 2004, but this update has not been published yet. Detailed descriptions of the variables included and classification according to the technological sectors to which the inventions patented belong (used also in the present paper) can be found in Hall, Jaffe, Trajtenberg (2002).

The number of citations itself does not indicate whether a patent is highly or lowly cited. Such information should be used comparatively. As stated by Hall, Jaffe, Trajtenberg (2002), determining the appropriate benchmark is complicated due to

* Universidad del País Vasco, Facultad de Ciencias Económicas, Bilbao, Spain (Petr.Mariel@ehu.es).
The authors wish to acknowledge financial support from the Department of Education of the Basque Government through grant IT-334-07 (UPV/EHU Econometrics Research Group), Ministry of Education and Science and FEDER (SEJ2005-05549/ECON, SEJ2007-61362)

DOI: 10.18267/j.pep.353

unavoidable time truncation of the number of citations given a start and end date of the data sample and due to time and technological area differences which also affect citation intensity. Determining how to treat those systematic differences in citation intensities is a challenging task.

Hall, Jaffe, Trajtenberg (2002) propose two different generic approaches. The first one, called the *fixed-effect approach*, proposes rescaling all citation intensities and expressing them as ratios to the mean citation intensity for patents in the same group of patents to which the patent in question belongs. This procedure eliminates any systematic changes over time, the truncation effect and effects caused by changes in the number of patents making citations. Unluckily many real effects can be lost because of this approach, as it does not attempt to separate real differences between groups of patents from those caused by truncation or propensity to cite.

The other approach, called the *quasi-structural approach*, is based on two identifying assumptions of proportionality and stationarity. The first of them states that the shape of the lag distribution over time is independent of the total number of citations received and the second that this distribution does not change over time. Our analysis is based on this approach.

Then, focusing on the benchmarking of patents, to the best of our knowledge there have been only parametric model based approaches to treating time and technological sector differences between patents, which is one of the points to be taken into account when comparing patent citations. In this paper we propose a more flexible solution to this problem by using nonparametric estimation techniques.

The rest of the paper is organized as follows. Section 2 describes the model and the estimation procedure applied based on Generalized Additive Models (GAM). Section 3 presents the outcomes obtained and compares them to already-published results based on the nonlinear estimation method. Finally, Section 4 concludes.

2. Model Specification and Estimation

We will focus our study on the analysis of the multiplicative citation model proposed by Hall, Jaffe, Trajtenberg (2001) and Hall, Jaffe, Trajtenberg (2002), which is based on the preliminary model studied in Caballero, Jaffe (2002) and Jaffe, Trajtenberg (2002*a*). The linearized logarithmic model equivalent to the multiplicative one proposed in the said papers is:

$$\begin{aligned} &\log(C_{k,s,t} / P_{k,s}) = \alpha_0 + \alpha_s + \alpha_t + \alpha_k + f_k(L) + u_{k,s,t} \\ &t = t_0, \ldots, T \\ &s = s_0, \ldots, t-1 \\ &k = 1, \ldots, K \end{aligned} \tag{1}$$

where $C_{k,s,t}$ is the total number of citations to patents in year s and technological field k coming from patents in year *t* and $P_{k,s}$ is the total patents observed in technological field k in year s. Hence the dependent variable represents the logarithm of the probability of citing an *(s,k)* patent in the citing year *t*. The coefficients $\alpha_j, j = 0, s, t, k$ are the logarithms of the coefficients from the initial multiplicative model. Coefficient α_0 is the common constant, the base, for a given field, cited-year and citing-year. The

DOI: 10.18267/j.pep.353

coefficients $\alpha_s, \alpha_t, \alpha_k$ correspond to the cited-year *(s)*, citing-year *(t)* and field *(k)* effects and must be interpreted relatively to the base group. That is, the value of $\exp(\alpha_k)$ indicates whether a patent belonging to technological field *k*, having the same cited-year and the citing-year as the patents considered in the base group, is more or less likely to be cited than those in the base group. Interpretations of α_s and α_t can be obtained in a similar way. The unknown function $f_k L$ depends on the citation-lag distribution ($L=t-s$) and $u_{k,s,t}$ is the error term.

Estimation of model (1) using parametric estimation methods is unfeasible without specifying the function that relates the citation-lag distribution ($f_k(L)$) with the lag (L) for each technological field (k). Since there is no information about these functions, Hall, Jaffe, Trajtenberg (2002) choose the same specification as Caballero, Jaffe (2002), who based their selection on rational features. They assume that each function $f_k(L)$ must be able to combine the effects of obsolescence and the diffusion path. The first characteristic, obsolescence, does not depend exactly on the course of time but is formed due to the accumulation of new ideas (patents) over time. The second, diffusion lag, appears because new ideas (patents) need some time to be seen by other inventors. Given this reasoning, the following function for describing the shape of the citation-lag distribution is proposed

$$f_k(L) = \exp(-\beta_{1k}L)(1-\exp(\beta_{2k}L)) \tag{2}$$

where β_{1k} represents the rate of obsolescence of knowledge and β_{2k} measures the diffusion rate. Despite the proposition of the same double exponential function for all technological fields, both rates are allowed to vary from one field to another. The ratio $1/\beta_{1k}$ becomes the modal lag, the lag that receives the highest citation frequency, and consequently the function $f_k(L)$ shifts to the left with larger values of β_{1k} . That is, the maximum is achieved earlier, so the obsolescence begins also earlier. The ratio β_{2k}/β_{1k} represents the highest citation frequency, so, increases on β_{2k}, *ceteris paribus*, result in higher citation intensities (see Caballero, Jaffe (2002) for more details).

This paper has two aims. First, we propose an alternative method for estimating (1) based on nonparametric techniques, which allows more flexible specifications of the citation-lag distribution function. Its flexibility is due to the possibility of estimating the citation-lag distribution function without imposing any structure, that is, we do not need to specify anything about how it varies over the lag. This helps to avoid the well known inconsistency problems derived from misspecification if the prior assumptions about specified function are incorrect. *A posteriori*, the function estimated through nonparametric techniques can be analysed to verify whether it contains both obsolescence and lag diffusion effects, and whether it follows a double exponential function.

The second objective of this paper is to generalize the specification of the citation model, allowing for more variability over the coefficients. This generalized model includes model (1) as a particular case and it reduces once again by introducing more variability problems of misspecification and offers the possibility of analyzing whether the cited and citing year coefficients are common for all technological fields.

The first objective is achieved by estimating model (1) using GAM. This estimation procedure, introduced by Hastie, Tibshirani (1990), is based on the simple idea that a dependent variable can be explained by the sum of a finite number of unknown functions that depend on one or more explanatory variables. There are several advantages of

DOI: 10.18267/j.pep.353

using this nonparametric estimation technique. First, because of its additive relation, it does not suffer from the problem derived from dimensionality, which is one of the few common problems in nonparametric estimation methods (see Härdle, 1990; Eubank, 1988). Second, it does not need to specify the unknown functions of the systematic part and the only assumption to be made is that these functions are smooth enough over their corresponding variables. Since each function can be any linear or nonlinear function, the classical linear regression model can be obtained when all functions are equal to a coefficient multiplied by the explanatory variable. In our case, model (1) brings together a combination of known and unknown functions. The functions for the field, cited and citing effects are considered as step functions (dummy variables) and the unknown functions correspond to the lag distribution functions for each field.

The GAM estimation procedure is based on the minimization of the sum of squared residuals subject to a penalizing term. Hastie and Tibshirani (1990) proved that using the backfitting algorithm consisting of minimizing the penalized sum of squared residuals converges to the unique solution of the system independently of the initial values. The flexibility of this method is so great that each function can be estimated by using different smoothers: Kernel methods, KNN, splines, local polynomials, *etc.* If the unknown functions are estimated by smoothing splines, the optimization problem for the estimation of model (1) is given by the following penalized sum of squared residuals:

$$\min_{\alpha_0,\alpha_t,\alpha_s,\alpha_k,f_k}\left\{\sum_{t=t_0+1}^{T}\sum_{s=s_0+1}^{S}\sum_{k=2}^{K}(\log(C_{k,s,t}/P_{k,s})-\alpha_0-\alpha_s-\alpha_t-\alpha_k-f_k(L))^2+\sum_{k=1}^{K}\lambda_k\int(f_k''(\ell))^2\,d\ell\right\} \quad (3)$$

where α_0 represents the base, so that for identification of all coefficients first technological field, citing-year and cited-year are dropped out from the first three sums. The function $f_k''(\cdot)$ stands for the second order derivative and λ_k, called the bandwidth or smoothness parameter, regulates the amount of smoothness imposed over its corresponding function. If λ_k is small, roughness is penalized lightly and consequently the amount of smoothness is low. In the limit case of no smoothness ($\lambda_k = 0, \forall k$) when roughness is not penalized at all, the optimization problem achieves its minimum with a null sum of squared residuals, a null bias for $\hat{f}_k(L)$ and an extremely high variance. In this case the resulting estimation is a perfect fit to the data and conveys no model information. Alternatively, as λ_k becomes larger, roughness is penalized more and the degree of smoothness imposed increases. In this case, the bias for $\hat{f}_k(L)$ increases but its variance tends to decrease. Finally, in the limit case of total smoothness $(\lambda_k \to \infty, \forall k)$ the estimated functions become linear, and the classical linear regression model is reached as a particular case. Therefore, the role of the smoothness parameters is to reach a trade-off between the asymptotic squared bias and error variance.

Hall, Jaffe and Trajtenberg (2002) estimate model (1) using minimization problem (3) without the penalizing term by nonlinear methods and substituting each function $f_k(L)$ by the double exponential function defined in (2) subject to the constraint $\sum_{k=1}^{K}\exp(f_k(L) = 1$. Given the parametric structure chosen for the lag distribution, not

DOI: 10.18267/j.pep.353

all coefficients are identified so an additional restriction is needed. Instead of allowing the effect of the cited-years (α_s) to vary over all years, its variability is restricted to five year intervals and the first two year interval is considered as the base.

As a first approach, we maintain the parametric part of model (1) almost unchanged. We merely allow the cited-year effect to vary over the whole range of years, and our key modification is introduced through the lag distribution functions since we do not impose any structure or restriction on them. Thus, we propose estimating model (1) by minimizing (3) using GAM methodology, not only because it offers a consistent estimator for all coefficients, including the lag distribution functions, but also because the estimation outcomes stand for a first descriptive exercise of issues inherent in the data. The nonparametric estimation method described is flexible enough to include the lag distribution function defined in (2) and subsequently the results obtained by using this method should be better than or at least as good as those obtained in Hall, Jaffe & Trajtenberg (2002). In this sense it seems more appropriate to estimate the lag functions without imposing any structure. After that, the estimated function for each sector can be compared to the double exponential function in order to verify its appropriateness. Therefore the GAM estimation is a useful descriptive tool, since it can throw some light on the question of what parametric function can be proposed in order to avoid problems of misspecification and to check *a priori* suspects.

The second objective of our paper deals with the amount of variability allowed for the remaining coefficients of the model, that is, whether the restrictions imposed on the alpha-coefficients in (1) can be empirically supported by the data. Let us summarize the assumptions of model (1): First, each technological field is assumed to have a different effect, *ceteris paribus*, on the probability of being cited but each remains constant over time. Second, the citing- and the cited-year effects are considered independent among years and, finally, citing- and cited-year effects are assumed to be common across the technological fields.

The first assumption seems to be quite reasonable since the technological fields have been formed by aggregating classes with the same characteristics. If the fields have been correctly classified a step function differentiating each field will be suffice.[1]

Nonetheless, the last two assumptions are hard to defend. First, let us analyse the second assumption in detail. According to the specification of model (1), the citing-year effect is independent over years. There is no doubt that this effect changes over years, but it is also true that drastic changes are not expected. A more realistic specification is to consider a smooth path, that is a smooth function varying over all citing-years, that is $\alpha_t = g_1(t)$. The same can be said about the cited-year effect so a smooth function varying over the cited years $\alpha_s = g_2(s)$ is considered. Thus the dummy variables are substituted by smooth functions, and the following generalization of model (1) is obtained:

$$\log(C_{k,s,t} / P_{k,s}) = \alpha_0 + \alpha_s + g_1(s) + g_2(t) + f_k(L) + u_{k,s,t} \qquad (4)$$

1 On the other hand, the assumption that the technological effect is constant over time can be questioned. If the model incorporates a time tendency term (α_t), then it is not appropriate to introduce another term that varies over time because it would cause an identification problem. Further analysis could consist of checking whether these two effects are additive, that is the appropriateness of considering $\alpha_k = \alpha_k + \alpha_t$, but for now we consider this first assumption reasonable.

DOI: 10.18267/j.pep.353

which can be estimated by GAM minimizing:

$$\min_{\alpha_0,\alpha_k,g_1,g_2,f_k}\left\{\sum_{t=t_0+1}^{T}\sum_{s=s_0}^{t-1}\sum_{k=2}^{K}\left(\begin{array}{l}(\log(C_{k,s,t}/P_{k,s})-\alpha_0-\alpha_k-g_1(s)-g_2(t)-f_k(L))^2\\+\lambda_1^0\int(g_1''(\ell))^2d\ell+\lambda_2^0\int(g_2''(\ell))^2d\ell+\sum_{k=1}^{K}\lambda_k\int(f_k''(\ell))^2d\ell\end{array}\right)\right\} \quad (5)$$

where in order to penalize roughness over the citing- and cited- year coefficients two penalty terms, controlled by the smoothing parameters λ_1^0 and λ_2^0, have been added. Large values of λ_1^0 and λ_2^0 imply a large amount of imposed smoothness and small differences in adjacent year effects. The smoothness introduced over the citing- and cited-year effects in the GAM estimation takes into account values of previous and subsequent years and not only observations corresponding to the same year as when dummy variables are used. If the control parameters λ_1^0 and λ_2^0 are equal to zero no smoothness is introduced, the functions $g_1(s)$ and $g_2(t)$ are estimated using a subsample containing observations corresponding to the same cited and citing years respectively. Subsequently, we reach the same results which are obtained by dummy variable specification (Model (1)).

Finally, when considering the third implicit assumption, there is no previous information that leads to restricting the citing and cited year effects to being the same across the different technological fields. In this sense, a natural generalization consists of allowing these effects to vary across defined industrial sectors, that is[2]

$$\log(C_{k,s,t}/P_{k,s})=\alpha_k^*+g_{1k}(s)+g_{2k}(t)+f_k(L)+u_{k,s,t}\,. \quad (6)$$

The functions $g_{1k}s$ and $g_{2k}(s)$ $g_{2k}(t)$ have the same meaning as in (4) but correspond to the k-th technological field. Model (6) can be estimated by GAM minimizing the following penalized sum of squared residuals:

$$\min_{\alpha_k^*,g_{1k},g_{2k},f_k}\left\{\sum_{t=t_0}^{T}\sum_{s=s_0}^{t-1}\sum_{k=1}^{K}\left(\begin{array}{l}(\log(C_{k,s,t}/P_{k,s})-\alpha_k^*-g_{1k}(s)-g_{2k}(t)-f_k(L))^2\\+\sum_{k=1}^{K}\lambda_{1k}^0\int(g_{1k}''(\ell))^2d\ell+\lambda_{2k}^0\int(g_{2k}''(\ell))^2d\ell+\lambda_k\int(f_k''(\ell))^2d\ell\end{array}\right)\right\}. \quad (7)$$

The restricted case of common citing-year and cited-year effects across technological fields leads to the same smoothness parameters for all technological fields, that is: $\lambda_{1k}^0=\lambda_1^0$ and $\lambda_{2k}^0=\lambda_2^0,\forall k$.

From this last generalized model, models (1) and (4) can be obtained as particular cases for determined smoothness degrees. In this sense, the selection of the smoothness parameters is an important prior step that must be tackled before minimizing the penalized sum of squared residuals. In order to select these smoothing parameters, several data driven methods have been proposed and analyzed in nonparametric estimation literature.

2 The relation between the alpha-coefficients is that $\alpha_k^*=\alpha_0+\alpha_k$ except for the base technological field for which $\alpha_k^*=\alpha_0$.

DOI: 10.18267/j.pep.353

3. Empirical Results

In this section we present parametric and semiparametric estimations of models (1), (4) and (6). First, we describe the results obtained by Hall, Jaffe, Trajtenberg (2002) and Hall, Jaffe, Trajtenberg (2001) where model (1) is estimated using nonlinear methods in order to solve the problem of truncated citations with the aim of constructing a citation-weighted stock of patents held by a firm. Therefore, by observing a given portion of the citation life, one could straightforwardly estimate the total citations of any specific patent. This estimation could be done by dividing the observed citations by the fraction of the population distribution that lies in the time interval for which citations are observed. Their results show that there are significant technological field effects and that the citing-year effect is clearly significant, presenting an increasing trend. By contrast, the cited-year effect is less variable and shows no clear pattern. The estimations of the parameters β_1 and β_2 define the citation-lag distribution by field after removing cited-year and citing-year effects. According their estimation results, citations in the Computers and Communications field appear as the fastest and citations in the Drugs and Medical field the slowest.

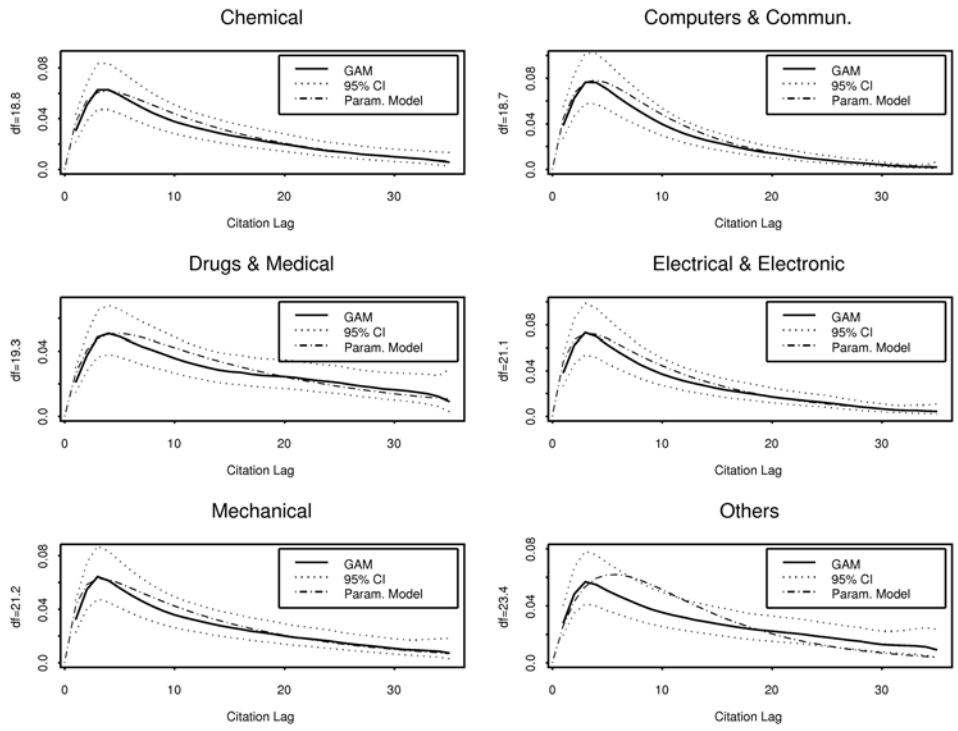
Our data set is the same as that used in Hall, Jaffe & Trajtenberg (2002) and aggregates records from six technological fields (Chemical, Computers and Communications, Drugs and Medical, Electrical and Electronics, Mechanical and Others). The application years run from 1963 to 1999 and the citation years from 1976 to 1999. Hall, Jaffe & Trajtenberg (2002) updates through 1999 the estimations presented in previous paper (Hall, Jaffe & Trajtenberg (2001)) based on data running from 1963 to 1994. [3].

Figures 1 and 2 present the estimation of model (1) using two different approaches. The first one, used by Hall, Jaffe & Trajtenberg (2002), Table 6 and labelled as Param. Model in the legend, assumes that the functions $f_k(L)$ are defined as double exponential functions of two parameters (see equation (2)) and the model is estimated by nonlinear least squares. The second one (labelled as GAM in the legend) leaves these functions unspecified and estimates them through smoothing splines (see minimization function (3)).

3 The added years, however, suffer from missing observations due to the truncation problem, which affects any approach. As the time gets closer to the final year, there is a considerable lack of patents filed in the last years that have not been granted yet and are therefore not included in the data set. Thus the estimated values for these years should be interpreted with caution

DOI: 10.18267/j.pep.353

Figure 1
Fraction of 35-Years Citation Total

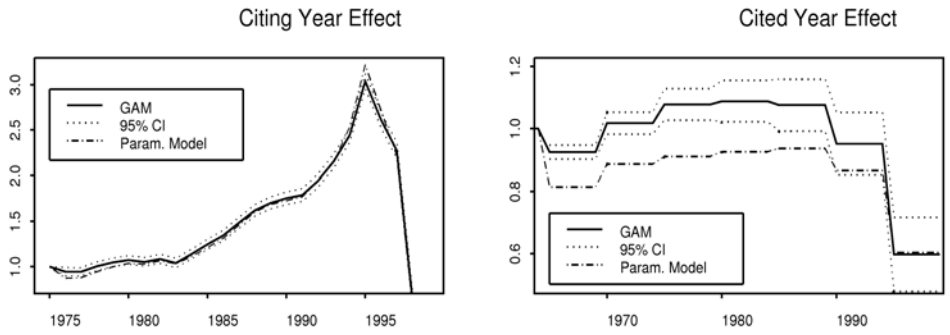


GAM: Parametric (Common Citing- and Cited-year effects) + Nonparametric (Sector dependent Citation lag)

Figure 1 shows the estimated citation-lag distribution under the two criteria together with the 95% confidence interval from the GAM estimation. The estimated citation-lag distributions are normalized to unity over 35 years as in Hall, Jaffe, Trajtenberg (2002) in order to be comparable. As can be observed from the figure, for all technological fields the specification of $f_k(L)$ as double exponential functions proposed by Caballero & Jaffe (2002) for model (1) belongs to the 95% confidence interval of the GAM estimation. Thus, the parametric functions they propose for the citation-lag distribution functions are supported by the data. Figure 2, on the other hand, shows the estimated citing- and cited-year effects and the corresponding 95% confidence interval of the GAM estimation. Taking into account these results, there are no significant differences between these two approaches in the citing and cited years (α_s and α_t in model (1)) when both are treated through dummy variables and only the non-linear part of the model, that is the citation-lag distribution, is treated differently.

DOI: 10.18267/j.pep.353

Figure 2
Citing and Cited Year Effects



GAM: Nonparametric (Sector dependent Citation lag)

Figures 3 and 4 compare the parametric estimation results of Model (1) using the exponential citation-lag distribution and the nonparametric estimation results obtained from Model (4). The differences between these two models rest not only on the specification of the citation-lag distribution, as before, but also on more flexibility in the specifications of the citing and cited effects. Notwithstanding the fact that these effects are still required to be the same for all technological fields, smooth time functions are considered for them, instead of treating them through dummies. As expected, the major differences appear in the shapes of these two effects.

Finally, Figure 5 (see the Appendix) compares the estimation results obtained from models (1) and (6). This last model is the most flexible approach analysed here since it allows time citing- and cited-year effects to be different across the technological fields. Given this important generalization, significant differences between the two approaches can be found. The major differences can be observed in the citing-year effect, which in the GAM approach presents a more stable behaviour. Regarding the citation-lag distribution estimation, the most important difference between the two estimations is found for the Computers & Communication sector, where the citation-lag distribution estimated by GAM presents a lower peak for the first years at the expense of its fatter tail.

Table 1 presents the residual sum of squares for the parametric model (see Hall, Jaffe, Trajtenberg (2002), Table 6) and the three models estimated by the nonparametric method presented above. It can be concluded that as well as introducing more flexibility into the original dummy-variable-based model, smaller residual sums of squares are obtained.

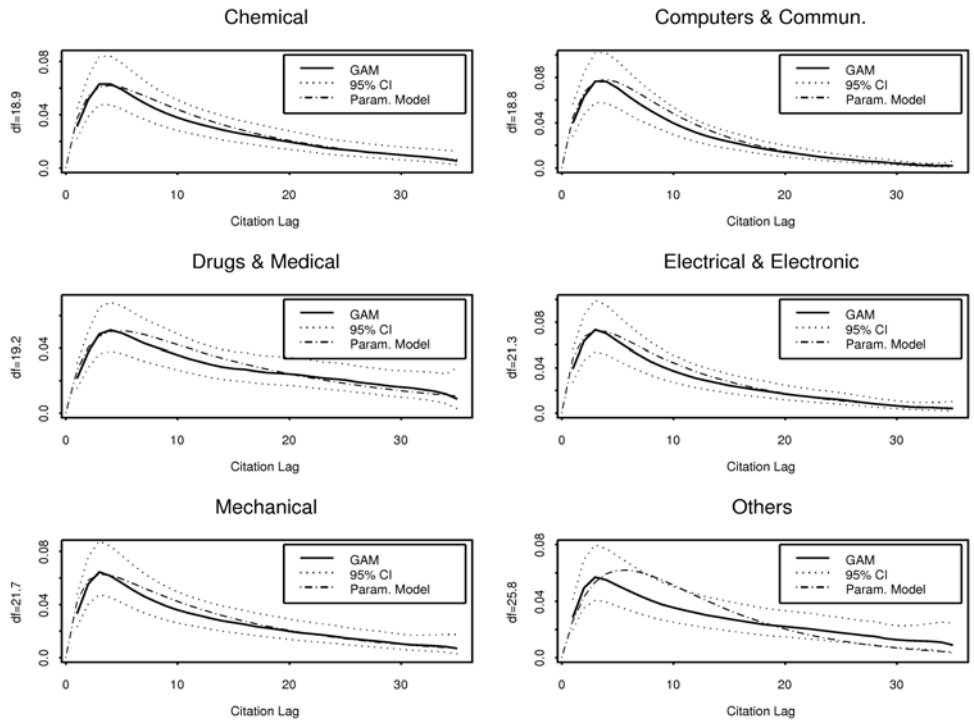
Table 1
Residual Sum of Squar

	Param. Model	**GAM**		
	Model (1)	**Model (1)**	**Model (4)**	**Model (6)**
RSS	112.49	73.71	71.79	57.05

DOI: 10.18267/j.pep.353

Let us focus now on the pure *propensity-to-cite* effect (Hall, Jaffe, Trajtenberg (2002), Table 7) in order to understand the importance of treating each field separately. The citing-year effect estimated is decomposed into the rise in the number of citing patents and the pure propensity-to-cite, which is obtained just by dividing the estimated citing-year function $g_{2k}(t)$ by the index of the number of potential citing patents by application year of the corresponding sector.

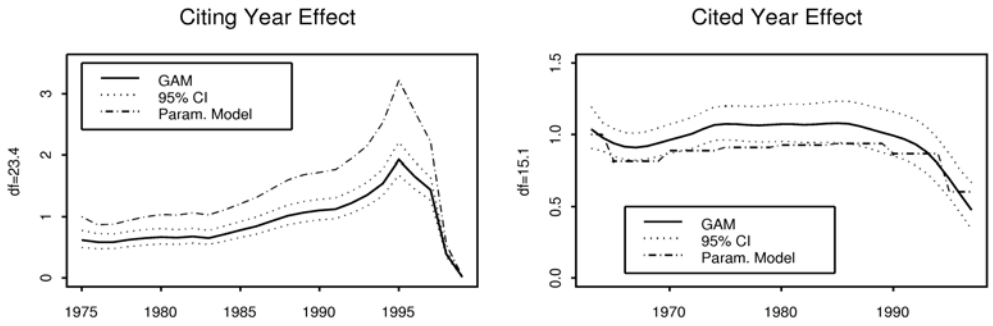
Figure 3
Fraction of 35-Years Citation Total



GAM: Nonparametric (Common Citing- and Cited-year effects, Sector dependent Citation lag)

DOI: 10.18267/j.pep.353

Figure 4
Citing- and Cited-Year Effects



GAM: Nonparametric (Common Citing- and Cited-year effects, Sector dependent Citation lag)

Figure 6 (see the Appendix) compares the pure propensity-to-cite effect. The common effect for all sectors obtained from the parametric model is represented by the solid line. The effects estimated for each field using GAM methodology are in dashed or dotted lines.

The general conclusion that can be drawn from the parametric model is that the pure propensity-to-cite increases until 1995 and that it accounts for about a 50% increase in citations. This conclusion is still valid for some but not all of the sectors under study when using the GAM methodology. Taking into account the differences between sectors, we can say that for sectors including Chemical, Mechanical and Others the pure propensity-to-cite is rising until 1995 by approximately 50%, as stated by the parametric model, but for the Electrical sector the increase is only about 30%. For the Drugs & Medical and Computers & Communication sectors this propensity has a decreasing trend for almost the whole sample period.

DOI: 10.18267/j.pep.353

Table 2
Total Patent Applications by Sector

	Chemical	Computers & Communi-cation	Drugs & Medical	Electrical & Electron-ics	Mechanical	Other
1975	15 436	4 129	3 807	10 361	16 924	15 231
1976	14 970	4 201	3 681	10 352	16 825	15 775
1977	15 063	4 237	3 775	10 387	16 568	15 948
1978	14 955	4 420	3 746	10 610	16 154	15 716
1979	14 675	4 651	4 143	10 543	16 252	15 462
1980	15 067	5 328	4 194	11 119	15 826	14 957
1981	14 466	5 458	4 313	10 693	14 843	14 137
1982	14 570	5 862	4 538	11 028	15 070	13 941
1983	13 395	5 750	4 408	10 428	14 038	13 544
1984	14 346	6 067	5 074	11 300	15 420	14 864
1985	14 852	6 626	5 706	12 125	16 694	15 439
1986	14 884	7 332	6 116	12 918	17 400	16 438
1987	16 112	8 334	6 990	13 865	18 345	17 812
1988	17 724	9 763	7 654	15 675	20 211	19 107
1989	18 961	10 910	8 166	17 129	21 005	19 906
1990	19 140	11 707	8 820	17 500	21 578	20 509
1991	18 800	12 899	8 805	18 108	21 544	19 860
1992	19 490	13 701	9 994	18 668	21 310	20 144
1993	19 407	14 872	11 427	18 791	21 363	20 988
1994	20 347	19 377	14 345	21 433	22 637	22 241
1995	23 496	24 602	18 659	23 306	24 147	23 451
1975-1995						
Increase	**52%**	**496%**	**390%**	**125%**	**43%**	**54%**

The next interesting step is a comparison between the common estimated pure propensity-to-cite effect and the raw change in the average number of citations per patent. Figure 7 (see the Appendix) shows these values, where averages present an increasing trend. For the common pure propensity-to-cite estimated by the parametric model there is an increase of about 5 to 10 between 1975 and 1995 (100%). It is observed that half of this increase is due to a rising pure propensity-to-cite but the other half is due to the higher number of patents available to be cited (see Hall, Jaffe, Trajtenberg, 2002, p. 446). However, if this comparison is made for each sector an interesting conclusion can be drawn. Figure 7 shows that the average number of citations *per* sector rises between 1975 and 1995 by about 100% in all sectors, but the pure propensity-to-cite effect presented in Figure 6 behaves differently for each sector. So following the previous reasoning, we can say that for the Drugs & Medical and Computers & Communication sectors (fast developing sectors for the years analysed) the increase in

DOI: 10.18267/j.pep.353

the average number of citations per patent is because there are more patents available to be cited and not because of a rising propensity-to-cite. This conclusion is also confirmed in Table 2, where the increases in the number of patents applications for the years mentioned in these two sectors are 390% and 496% respectively, approximately eight and ten times higher than for the Chemical, Mechanical and Other sectors. The situation is similar for the Electrical & Electronics sector, where the increase in the number of applied patents is 125% and pure propensity-to-cite increases more slowly than in the Chemical, Mechanical and Other sectors for which the numbers of patent applications only rise 52%, 43% and 54%, respectively.

4. Conclusion

The estimation of the citation model using GAM methodology allowing for different behaviours of the cited- and citing-year effect for each technological sector brings new knowledge about the NBER Patent-Citations Data File. This methodology generalizes the parametric model estimated in Hall, Jaffe, Trajtenberg (2002), but for specific values of smoothness degrees, these results can be obtained as a particular case. As mentioned above, results of the nonparametric methods can be employed as a descriptive tool. Regarding the lag-distribution, and answering the first question of our paper, we conclude that the double exponential function defined in Hall, Jaffe, Trajtenberg (2002) is a reasonable proposal supported by the data. In answer to the second question of our paper, important differences between sectors are obtained for the citing- and cited-year effects. Thus the requirement that these effects must be common to all sectors is not supported by the data and its imposition leads to misspecification and incorrect conclusions.

Hall, Jaffe, Trajtenberg (2002), p. 415 describe important changes in the shares of total patent applications over time according to the six technological categories, and find a steady decline in the three traditional fields (Chemical, Mechanical, and Others), stable behaviour for Electrical & Electronics and steep increases for the Computers & Communication and Drugs & Medical fields (see Table 2). They state that "*this reflects the much-heralded technological revolution of our times, associated with the rise of information technologies and the growing importance of health care technologies*". We add to this conclusion that the pure propensity-to-cite proper to each technological field is closely related to these shares and it highlights a clear difference between traditional and advanced fields.

Treating all technological sectors together could lead to the erroneous conclusion that the common rising propensity-to-cite is a reflection of higher fertility for more recent cohorts of patents or for an artifactual change in the propensity. Nevertheless, the decreasing propensity to cite accompanied by an increasing number of citations in advanced sectors invalidates this reasoning, showing that the only cause in these fields is the high fertility of recent cohorts. Nonetheless, we do not know whether the differences between sectors in citations reflect a real phenomenon or different citation practices that are artifactual. They are probably a result of a mixture of the two causes and should be analysed in future research.

DOI: 10.18267/j.pep.353

References

Basberg, B. (1987), 'Patents and the Measurement of Technological Change: A Survey of the Literature', *Research Policy* 16, 131–141.

Caballero, R. J., Jaffe, A. B. (2002), How High Are Giants' Shoulders: An Empirical Assessment of Knowledge Spillovers and Creative Destruction in a Model of Economic Growth, *in* A. B. Jaffe & M. Trajtenberg, eds, 'Patents, Citations, and Innovations. A Widow on the Knowledge Economy', The MIT Press, Cambridge, Massachusetts, London, pp. 89-152.

Eubank, R. L. (1988), *Spline Smoothing and Nonparametric Regression*. New York: Marcel Dekker, Inc.

Griliches, Z. (1990), 'Patent Statistics as Economic Indicators: A survey', *Journal of Economic Literature* 28, pp. 1661-1707.

Hall, B. H., Jaffe, A., Trajtenberg, M. (2001), Market Value and Patent Citations: A First Look, University of California, Berkeley. Working Paper No. E01-304.

Hall, B. H., Jaffe, A., Trajtenberg, M. (2002), The NBER Patent-Citations Data File: Lessons, Insights, and Methodological Tools, *in* A. B. Jaffe & M. Trajtenberg, *eds,* 'Patents, Citations, and Innovations. A Widow on the Knowledge Economy', The MIT Press, Cambridge, Massachusetts, London, pp. 403–459.

Hall, B. H., Jaffe, A., Trajtenberg, M. (2005), 'Market Value and Patent Citations', *RAND Journal of Economics* 36, pp. 16–38.

Härdle, W. (1990), *Applied Nonparametric Regression*. New York: Cambridge University Press,.

Hastie, T. J., Tibshirani, R. J. (1990), *Generalized Additive Models*. London: Chapman and Hall.

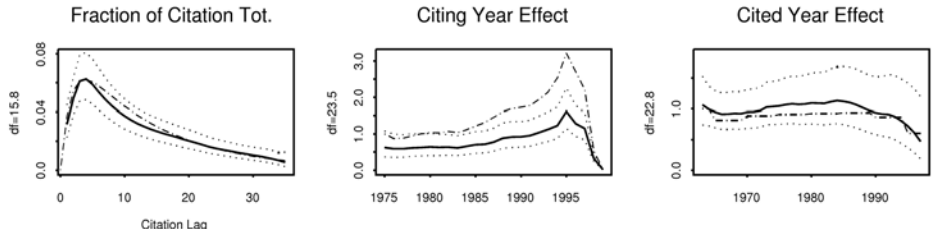
Jaffe, A. B., Trajtenberg, M. (2002*a*), International Knowledge Flows: Evidence from Patent Citations, *in* A. B. Jaffe & M. Trajtenberg, *eds*, 'Patents, Citations, and Innovations. A Widow on the Knowledge Economy.' Cambridge (Massachusetts), London: The MIT Press, pp. 199-234.

Jaffe, A. B., Trajtenberg, M. (2002*b*), *Patents, Citations, and Innovations. A Widow on the Knowledge Economy*. Cambridge (Massachusetts), London: The MIT Press.

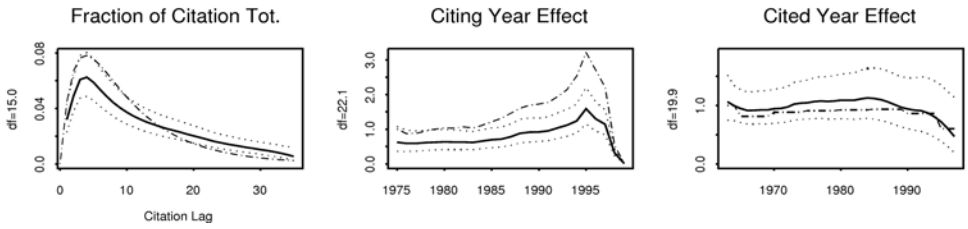
APPENDIX

Figure 5

Parametric and Nonparametric Estimation Results for Models (1) and (6)

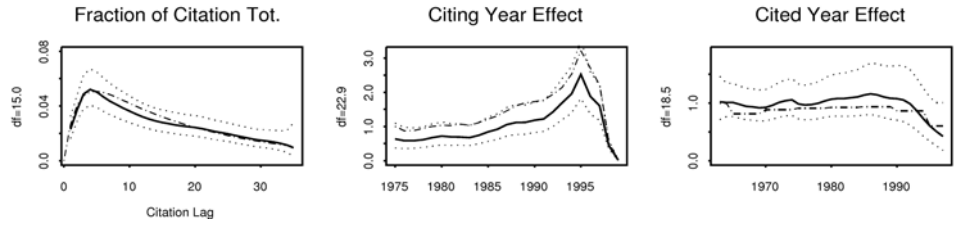


Computers & Communications

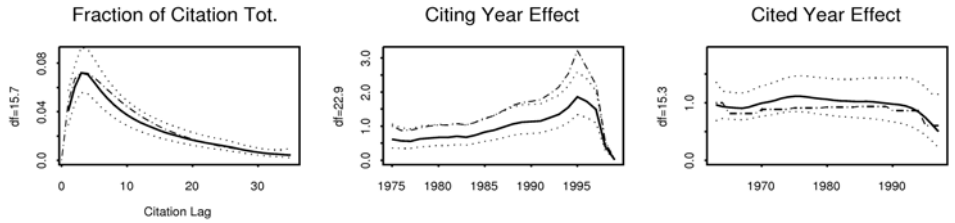


DOI: 10.18267/j.pep.353

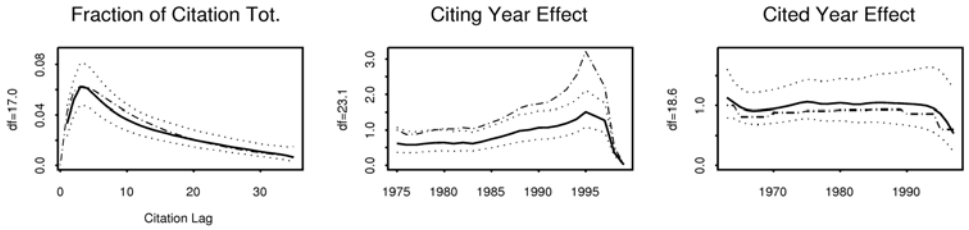
Drugs & Medical



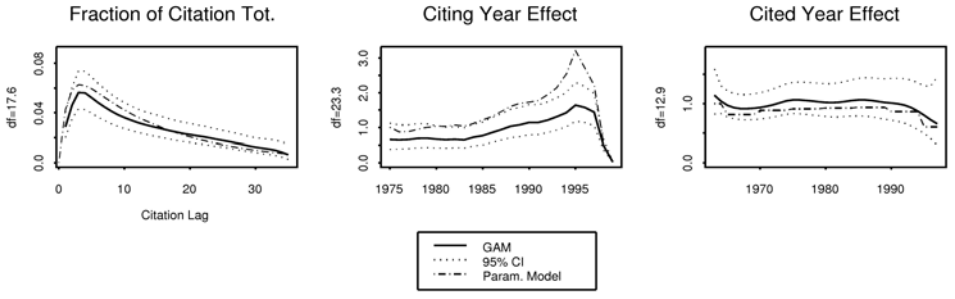
Electrical & Electronics



Mechanical



Others



DOI: 10.18267/j.pep.353

Figure 6
Pure Propensity to Cite Effect

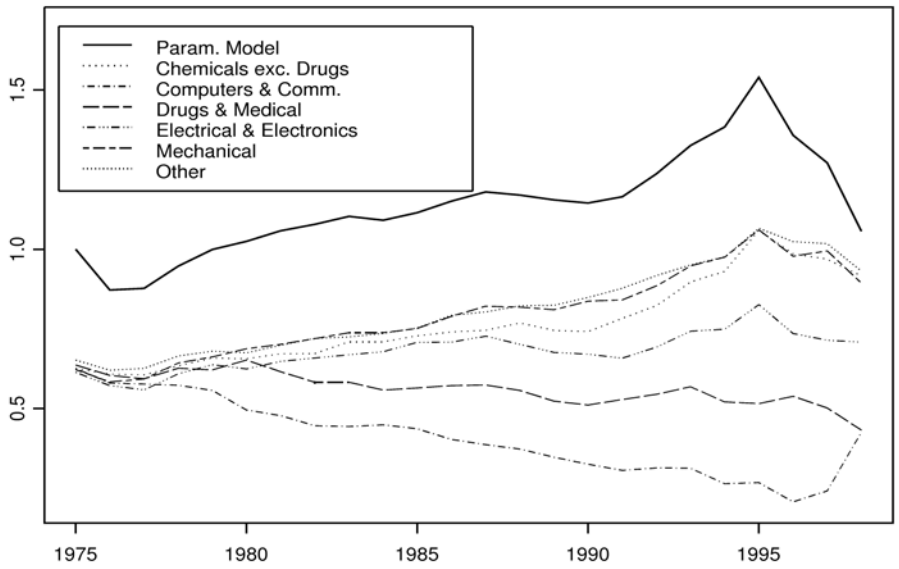


Figure 7
Average Number of Citations Made by Sectors