# Multiple regression analysis of a patent's citation frequency and quantitative characteristics: the case of Japanese patents

Fuyuki Yoshikane*

*Graduate School of Library, Information and Media Studies, University of Tsukuba, 1-2, Kasuga, Tsukuba, Ibaraki 305-8550, Japan*

e-mail: fuyuki@slis.tsukuba.ac.jp; Tel.: +81-29-859-1346

* Corresponding author

**Abstract** Although many studies have been conducted to clarify the factors that affect the citation frequency of "academic papers," there are few studies where the citation frequency of "patents" has been predicted on the basis of statistical analysis, such as regression analysis. Assuming that a patent based on a variety of technological bases tends to be an important patent that is cited more often, this study examines the influence of the number of cited patents' classifications and compares it with other factors, such as the numbers of inventors, classifications, pages, and claims. Multiple linear, logistic, and zero-inflated negative binomial regression analyses using these factors are performed. Significant positive correlations between the number of classifications of cited patents and the citation frequency are observed for all the models. Moreover, the multiple regression analyses demonstrate that the number of classifications of cited patents contributes more to the regression than do other factors. This implies that, if confounding between factors is taken into account, it is the diversity of classifications assigned to backward citations that more largely influences the number of forward citations.

**Keywords** Patent citation; Citation frequency; Regression analysis; Japan

# Introduction

The citation frequency of academic papers is considered to express their importance in some sense, and therefore it is often used as a measure in research evaluation. In recent years, the same type of study has been conducted for patents. Of course, similar to the case of citations between papers,[1] there are not only positive citations but also negative ones between patents, and their objectives and levels of indispensability vary. However, patent citations are basically perceived as the reutilization of an existing technology (Inuzuka 2011), and it is thought that a citation denotes the citing patent has found utility value in the cited patent. In fact, some studies reported relationships between the importance or economic value of a patent and its citation frequency (e.g., Narin 1995; Harhoff et al. 1999). In other words, to predict the citation frequency of patents or to grasp factors affecting it is meaningful in that it helps us to estimate the importance of patents.

Many studies have been conducted to understand the factors that affect the citation frequency of academic papers on the basis of statistical analysis such as multiple regression analysis (e.g., Peters and van Raan 1994; Bornmann and Daniel 2007). On the other hand, with regard to patent documents, while several studies have attempted to distinguish important patents using citation frequency as one of the explanatory variables, there have been few studies in which the citation frequency has been explained and predicted as the response variable.

Yoshikane et al. (2012) analyzed the relationship between the citation frequency of patents and the diversity of their citations (the number of different classifications associated with patents cited by them). They reported that (i) although the correlation between both was statistically significant, the values of the correlation coefficient were low at approximately 0.1 and that (ii) when patents were grouped by the citation frequency and the diversity of citations was compared between the groups, the diversity of citations in the often cited group was between 1.5 and 4 times as high as that in the less frequently cited group. These results indicate that although the two do not have a simple, linear correlation, there is a possibility that the number of times a patent will be cited (i.e., the number of forward citations) is affected by the diversity of the patents that were cited in it (i.e., the diversity of backward citations). However, they paid attention only to the diversity of the cited patents as a factor; they did not perform an analysis in which multiple factors are comprehensively considered. Therefore, we cannot deny the possibility that only an “apparent correlation,” where multiple factors are confounded, is shown.

With the above as background, in the present study, I perform multiple regression analyses

[1] Bornmann and Daniel (2008) introduced and reviewed citation categorization regarding academic papers from various viewpoints.

that explain the citation frequency with multiple factors assumed to influence it. In addition to the number of cited patents' classifications, I consider eight variables, such as the numbers of inventors, pages, and claims. This study aims to clarify the correlation between the citation frequency and the number of cited patents' classifications under the condition that other factors are controlled. Moreover, the influence of the number of cited patents' classifications on the citation frequency is compared with that of other factors based on the contribution of variables in the regression models.

This paper is organized as follows. First, I describe the source data used in the analyses. Then, after explicating the methodology, including the variables, I present results of the multiple regression analyses, in which the citation frequency is explained by variables representing the quantitative characteristics of patents. Finally, on the basis of these results, I discuss the influence of the factors on the citation frequency of patents.

# Data

Information sources in this study were the NTCIR test collections compiled by the National Institute of Informatics (NII), Japan, and I used the full text of the "patent gazette (publication of unexamined patent applications)" published in Japan; the 3,496,253 documents published in the ten years between 1993 and 2002 from NTCIR-7 Patent Mining Test Collection (Nanba et al. 2008); and the 1,757,361 documents published in the five years between 2003 and 2007 from NTCIR-8 Patent Translation Test Collection (Fujii et al. 2010). Approximately 350,000 documents were published in each of these years. A total of 341,388 patent applications published in 1998 were subjected to the analysis. I investigated the classifications of patents cited by them and the number of times each of them is cited among the ten years following their publication, that is, from 1998 to 2007. As for the classifications assigned to the cited patents, the investigation was based on the "patent gazette" published during the period in which the data were available, that is, 1993 onward. Therefore, if a patent published in 1992 or earlier was cited, its classifications cannot be identified. Despite this limitation, however, I consider it reasonable to assume that it does not have a serious effect on the results, which will reveal general tendencies of citation among patents.

There are cases in which the descriptions of cited patents are inserted in the main text of the patent document rather than as independent items. Moreover, the format of these descriptions is not standardized. Thus, it is difficult to completely and precisely extract the information of cited patents, particularly for the patents published before 2002 when the information disclosure system for prior art documents was introduced in Japan (Sato and Iwayama 2006). Furthermore, there are numerous instances of typographical errors that are assumed to have occurred with digitization of patent documents, such as mistakes in the software conversion of Kana (Japanese phonetic alphabets) to Kanji (ideographic characters) and those in the OCR conversion (Inuzuka 2011). Similar problems also exist in extracting or calculating some of the feature values of the patent application used in the multiple regression analyses as explanatory variables, which will be indicated in the next section. This study has searched for and listed variants of the patterns of description in patent documents through data observation, and has covered those variations in obtaining values of each variable.

# Methods

Assuming that a patent based on a variety of technological bases tends to be an important patent that is cited more often, this study examines the hypothesis that the number of cited patents' classifications is positively related to the citation frequency under the condition that other factors are controlled. For this purpose, multiple regression analyses using the following variables are executed. The influences on the citation frequency are compared between explanatory variables through multiple regression analyses.

## Response variable

The response variable is the number of times the subject patent is cited by others, namely its citation frequency, during the ten years after its publication ($F_{cited}$). Figure 1 shows the distribution of the citation age, which means how many years later a patent is cited after its publication (the difference between the cited and citing patents, not for one's publication and the other's application years but for the publication years of both). The distribution is expressed per "section," which is the top layer in the International Patent Classification (IPC). Section A is "human necessities," B is "performing operations; transporting," C is "chemistry; metallurgy," D is "textiles; paper," E is "fixed constructions," F is "mechanical engineering, etc.," G is "physics," and H is "electricity" (WIPO 2010). In general, the distribution of the citation age has similar tendencies for all sections. Regarding section D, the proportion of citations in the period just after publication (during a few years) is small compared to the other sections. For all sections, there are still many citations even after nine years. So, in order to get an overall picture of the citation behavior for patents, it would be necessary to observe data over a longer time period. However, after reaching a peak at around six years later, the frequency of being cited begins to decrease. I consider that covering ten years in the observation would provide at least an understanding in broad outline.

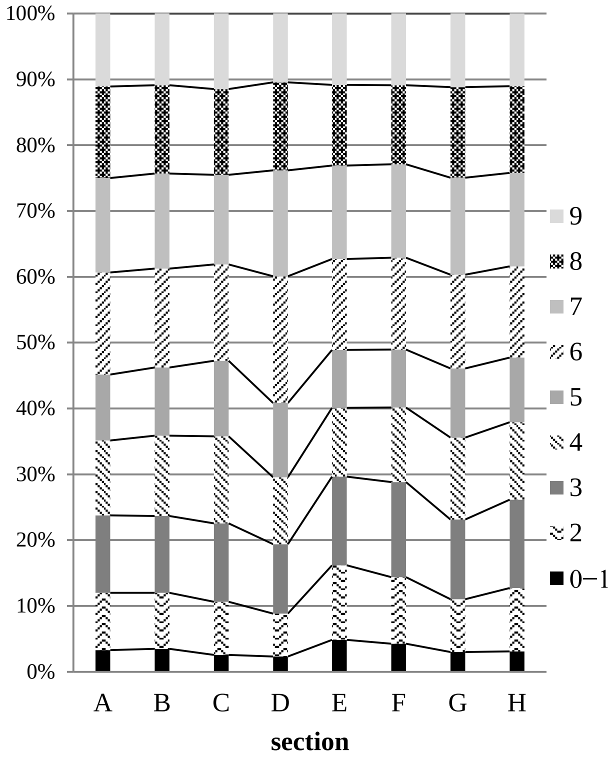


**Fig. 1** Citation age of patents for each section

## Explanatory variable

As for explanatory variables, I adopt the numbers of inventors (*IV*), associated classifications (*VC*), pages (*PG*), figures (*FG*), tables (*TB*), claims (*CL*), priority claims (*PC*), countries for priority claims ($PC_c$), and classifications associated with the patents that the subject patent is citing ($VC_{citing}$). Table 1 details the procedures employed to derive these indices from the data and the maximum value for each. In multiple regression analyses, explanatory variables other than $VC_{citing}$ are regarded as control variables.

The above indices are adopted because they are considered to influence the citation frequency. While the number of inventors for a patent application corresponds to the number of authors for an academic paper, the numbers of pages, figures, and tables represent quantities of descriptions in a document. These indices have been dealt with in the correlation or regression analyses of the citation frequency targeting academic papers (e.g., Snizek et al. 1991; Glänzel 2002; Kostoff 2007). Supposing that these correlate with the citation frequency for patent applications as well as for academic papers, this study uses *IV*, *PG*, *FG*, and *TB*. On the other hand, the numbers of claims *CL*, priority claims *PC*, and countries for priority claims $PC_c$ are quantities specific to patent applications. This study also supposes that these indices, which are related to the rights and value of inventions, correlate with the citation frequency. The number of classifications of the subject patent *VC* and the number of classifications of patents cited by it (its backward citations) $VC_{citing}$ are quantities that reflect the diversity of the invention's contents, and the relationship of these with the citation frequency of the subject patent—that is, the number of times it is cited by others (its forward citations)—has been pointed out (Yoshikane et al. 2012). According to Yoshikane et al. (2012), looking into classifications at the subclass level (the fourth layer in IPC), which is more detailed than the section level (the top layer in IPC), gives higher values of the correlation coefficient with the citation frequency. This study, therefore, counts the number of classifications at the subclass level.

Table 1 Procedure of derivation of each index

| Index | Procedure of Derivation | Maximum Value |
|---|---|---|
| Citation frequency ($F_{cited}$) | Counting the number of patents that refer to the subject patent's publication (or application) number within the field of the "prior art documents" or within the main text | 898 |
| Number of inventors (*IV*) | Counting the number of tags of "(72) Inventor" | 22 |
| Number of associated classifications (*VC*) | Extracting all classifications from the field of the "IPC," and then counting the number of different classifications at the subclass level | 24 |
| Number of pages (*PG*) | Extracting the value from the field of the "total number of pages" | 775 |
| Number of figures (*FG*) | Extracting the largest value assigned to the figure number in captions | 453 |
| Number of tables (*TB*) | Extracting the largest value assigned to the table number in captions | 616 |
| Number of claims (*CL*) | Extracting the value from the field of the "number of claims" or, for the case where this field is not present, extracting the largest value assigned to the claim number in captions | 358 |
| Number of priority claims (*PC*) | Counting the number of tags of "(31) Priority Claim Number" | 21 |
| Number of countries for priority claims ($PC_c$) | Extracting a country corresponding to each priority claim number and then counting the number of different countries | 3 |
| Number of classifications associated with backward citations ($VC_{citing}$) | Extracting all classifications for each patent whose publication (or application) number is referred to in the subject patent and then, as with *VC*, counting the number of different classifications at the subclass level | 25 |

The numbers enclosed by brackets in tags, such as the "72" in "(72) Inventor," are the INID (Internationally agreed Numbers for the Identification of bibliographic Data) codes.

Figure 2 is a box plot that shows for each index the distribution of values, which are normalized to [0, 1] by dividing the maximum value of that index. The right-hand side of each box shows the 75th percentile value. Regarding indices other than the numbers of inventors (*IV*) and associated classifications (*VC*), the values are concentrated in the area around each minimum value. Furthermore, we can confirm that these two (*IV* and *VC*) also have very skewed distributions, in which the 75th percentile value is no more than one-twentieth of the maximum value. Since all these indices have highly skewed distributions and do not follow a normal distribution, I transformed each index, *x*, into natural logarithmic values, ln(*x*+1), before calculating correlation coefficients between them and applying them as response/explanatory variables to the multiple regression analyses. I added one to the values for avoiding zero in the logarithmic transformation.

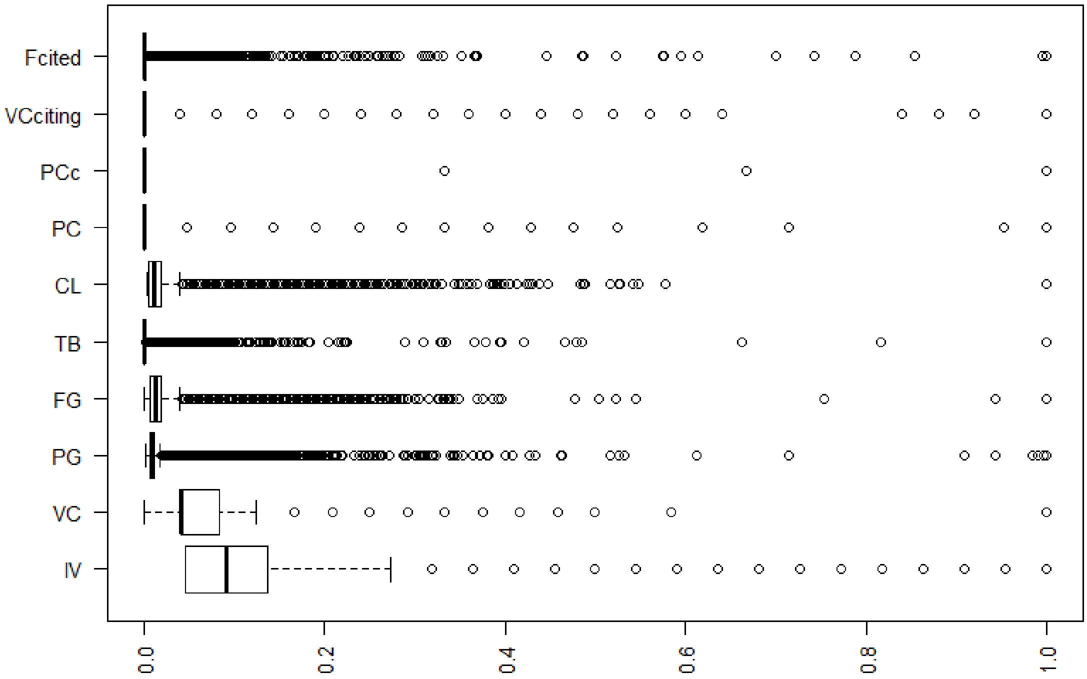


**Fig. 2** Distribution of normalized values for each index

First, I examine simple correlations between explanatory variables and between each of these and the response variable, i.e., $F_{cited}$. Pearson's product-moment correlation coefficient is calculated for each pair of variables on the basis of the values following the logarithmic transformation. Considering the multicollinearity between variables judged from the observed values of the correlation coefficients, I select variables to be excluded as the need arises. These variables are not included as explanatory variables in the following multiple regression analyses.

In most patent applications, the value of the citation frequency, which is used as the response variable, is zero, as stated above. This fact may make it difficult to successfully apply linear regression to the data. Thus, this study is based not only on a linear model but also on a logistic model, wherein patents whose citation frequency is equal to or beyond a certain threshold can be differentiated from others. Furthermore, zero-inflated negative binomial (ZINB) regression is applied. ZINB models, which are robust against overdispersion caused by a large number of zero counts, are used in bibliometric studies, including patent analyses (Foltz et al. 2000; Odagiri et al. 2002; Lee et al. 2007; Tang and Shapira 2012).

For the linear regression analysis, I introduce explanatory variables selected through the stepwise method, setting the variable inclusion criteria at the statistically significant probability value (*p*-value), $p_{in} < 0.05$, and the variable exclusion criteria at $p_{out} > 0.05$. For the logistic regression analysis, on the other hand, the following models are adopted.

(1) A model that includes all explanatory variables (except those excluded on the basis of simple correlations)

(2) A model that excludes one explanatory variable from the model in (1)

In the analysis of (2), each model that excludes each of the explanatory variables is adopted and examined in turn. To assess the contribution of these variables to the regression and infer the influence of factors on the number of citations, I observe fluctuations for the fitness of the regression model when each variable is excluded. Moreover, by changing the value of threshold *k*, which separates the often cited and less frequently cited patents, from 1 to 10, I perform different sorts of discrimination on the basis of the logistic regression, from “discriminating patents being cited at least once” to “discriminating those being cited very often” by degrees. Lastly, by applying ZINB regression, the number of cited patents’ classifications is reconfirmed to significantly correlate with the citation frequency.

# Results

## Simple correlations between indices

A total of 341,388 patents published in 1998 were subjected to the analysis. In respect to the whole of these, I calculated the product-moment correlation coefficient *r* for each pair among the ten variables: the response variable, i.e., $F_{cited}$ (citation frequency), and the nine explanatory variables, i.e., *IV* (inventors), *VC* (classifications), *PG* (pages), *FG* (figures), *TB* (tables), *CL* (claims), *PC* (priority claims), $PC_c$ (countries for priority claims), and $VC_{citing}$ (classifications associated with backward citations). Values of the correlation coefficient *r* are presented in Table 2.

Table 2 Correlations between variables

| | ***IV*** | ***VC*** | ***PG*** | ***FG*** | ***TB*** | ***CL*** | ***PC*** | $\boldsymbol{PC_c}$ | $\boldsymbol{VC_{citing}}$ | $\boldsymbol{F_{cited}}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| ***IV*** | - | 0.07 | 0.18 | −0.06 | 0.20 | 0.13 | 0.11 | 0.11 | 0.09 | 0.09 |
| ***VC*** | 0.07 | - | 0.09 | −0.12 | 0.14 | 0.09 | 0.06 | 0.05 | 0.11 | 0.07 |
| ***PG*** | 0.18 | 0.09 | - | 0.42 | 0.21 | 0.52 | 0.25 | 0.24 | 0.15 | 0.12 |
| ***FG*** | −0.06 | −0.12 | 0.42 | - | −0.41 | 0.22 | 0.00 | −0.01 | −0.07 | −0.03 |
| ***TB*** | 0.20 | 0.14 | 0.21 | −0.41 | - | 0.06 | 0.11 | 0.11 | 0.17 | 0.12 |
| ***CL*** | 0.13 | 0.09 | 0.52 | 0.22 | 0.06 | - | 0.35 | 0.35 | 0.05 | 0.10 |
| ***PC*** | 0.11 | 0.06 | 0.25 | 0.00 | 0.11 | 0.35 | - | 0.98 | −0.04 | 0.03 |
| $\boldsymbol{PC_c}$ | 0.11 | 0.05 | 0.24 | −0.01 | 0.11 | 0.35 | 0.98 | - | −0.04 | 0.02 |
| $\boldsymbol{VC_{citing}}$ | 0.09 | 0.11 | 0.15 | −0.07 | 0.17 | 0.05 | −0.04 | −0.04 | - | 0.11 |
| $\boldsymbol{F_{cited}}$ | 0.09 | 0.07 | 0.12 | −0.03 | 0.12 | 0.10 | 0.03 | 0.02 | 0.11 | - |

First, I discuss the correlations between explanatory variables. A strong correlation was observed only for the two variables relating to the priority claim, that is, *PC* and $PC_c$ (priority claims and countries for priority claims). The value of the correlation coefficient between the two was extremely high at 0.98. Rather strong positive correlations of more than 0.40 were seen between *PG* and *CL* (pages and claims) ($r$ = 0.52) and between *PG* and *FG* (pages and figures) ($r$ = 0.42). On the other hand, the pair of *FG* and *TB* (figures and tables) showed a rather strong negative correlation ($r$ = −0.41). This implies that there is a tendency for patents to be separated into two types: patents that explain the content of invention mainly through figures and those that explain it mainly through tables. Among the nine explanatory variables, only *PG* (pages) and *CL* (claims) demonstrated positive correlations with all variables. This may be due to the fact that they directly reflect the general volume of descriptions in patent applications.

Next, I discuss the correlations of the explanatory variables with the response variable, $F_{cited}$ (citation frequency). All explanatory variables except *FG* (figures) had positive values of the correlation coefficient with $F_{cited}$. In other words, patents with a large "amount" in terms of application contents or inventors tend to be frequently cited. However, correlations between the explanatory variables and $F_{cited}$ were generally weak. The absolute values of the correlation coefficient with $F_{cited}$ were less than 0.20. Compared with the other variables, *PG* (pages) and *TB* (tables) showed relatively strong correlations with $F_{cited}$. These were followed by $VC_{citing}$ (classifications associated with backward citations).

As mentioned previously, Yoshikane et al. (2012) indicated the relationship between the citation frequency, i.e., the number of forward citations, and the number of classifications assigned to cited patents, i.e., the diversity of backward citations. As with the results reported in Yoshikane et al. (2012), it is $VC_{citing}$ rather than *VC*—that is to say, the number of classifications of patents which a particular patent is citing rather than the patent itself—that has a stronger relationship with the citation frequency. However, in the comparison of simple correlations, the correlation coefficient of some variables (*PG* and *TB*) is slightly higher than that of $VC_{citing}$.

Among the nine explanatory variables, the pair of *PC* and $PC_c$ (priority claims and countries for priority claims) has an extremely high value of the correlation coefficient. Therefore, in light of the multicollinearity that would be caused by the two variables, I included only one of them, not both, as an explanatory variable in the multiple regression analyses presented below. Among the two, $PC_c$ (countries for priority claims) was excluded for the following two reasons: (1) $PC_c$ shows a slightly weaker correlation with the response variable $F_{cited}$ than does *PC*, and (2) because the number of countries for priority claims has a low maximum value of 3 (see Table 1) and does not widely vary with the patent, it is considered that $PC_c$ would not function very effectively as a feature value.

## Multiple linear regression for citation frequency

Table 3 shows results of the multiple linear regression analysis conducted for each section of A–H as well as for all patents published in 1998. The following values are presented in the table: the number of subject patents *n*, coefficient of determination adjusted for degrees of freedom $R'^2$, and standardized partial regression coefficient for each explanatory variable. Since it is common to assign multiple classifications to one patent, the sum of *n* under the classifications A–H is greater than the total *n* of patents (the bottom row of Table 3). Variables that were not selected by the stepwise method are expressed as "*N.S.*"

*PG* (pages) and $VC_{citing}$ (classifications associated with backward citations) had relatively high values of the standardized partial regression coefficient compared with the other six explanatory variables. In particular, regarding $VC_{citing}$, the regression coefficient was statistically significant ($p <$ 0.001) throughout all the sections and the values remained comparatively high. Regarding *PG*, on the other hand, the regression coefficient was not significant and its absolute value was very low in section A (human necessities). *VC* (classifications of the subject patent) also had significant positive values of the regression coefficient in all sections except section A. Nevertheless, it is not the number of classifications of a patent itself but that of the patents it is citing, i.e., $VC_{citing}$, that tends on the whole to have higher values of the regression coefficient. The result exhibiting that $F_{cited}$

(citation frequency) had a stronger relationship with $VC_{citing}$ than with *VC* was common to both the observation of simple correlations and the multiple regression analysis. *IV* (inventors), *TB* (tables), and *CL* (claims) also had higher values of the regression coefficient than had *VC*.

Table 3 Results of the multiple linear regression analysis

| | ***n*** | $R'^2$ | Standardized partial regression coefficient | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | ***IV*** | ***VC*** | ***PG*** | ***FG*** | ***TB*** | ***CL*** | ***PC*** | $VC_{citing}$ |
| A | 39474 | 0.035$^*$ | 0.046$^*$ | *N.S.* | 0.016 | −0.038$^*$ | 0.097$^*$ | 0.060$^*$ | −0.063$^*$ | 0.071$^*$ |
| B | 92483 | 0.048$^*$ | 0.046$^*$ | 0.051$^*$ | 0.099$^*$ | −0.056$^*$ | 0.059$^*$ | 0.044$^*$ | *N.S.* | 0.078$^*$ |
| C | 46881 | 0.029$^*$ | 0.047$^*$ | 0.029$^*$ | 0.049$^*$ | −0.021$^*$ | 0.043$^*$ | 0.043$^*$ | −0.015 | 0.099$^*$ |
| D | 6255 | 0.073$^*$ | 0.070$^*$ | 0.070$^*$ | 0.067$^*$ | −0.121$^*$ | 0.072$^*$ | 0.030 | −0.037 | 0.069$^*$ |
| E | 22564 | 0.015$^*$ | 0.036$^*$ | 0.041$^*$ | 0.030$^*$ | *N.S.* | 0.038$^*$ | 0.053$^*$ | *N.S.* | 0.031$^*$ |
| F | 40437 | 0.029$^*$ | 0.049$^*$ | 0.050$^*$ | 0.083$^*$ | 0.014 | 0.034$^*$ | 0.027$^*$ | −0.024$^*$ | 0.064$^*$ |
| G | 103919 | 0.042$^*$ | 0.049$^*$ | 0.048$^*$ | 0.085$^*$ | −0.042$^*$ | 0.047$^*$ | 0.077$^*$ | −0.022$^*$ | 0.086$^*$ |
| H | 96679 | 0.036$^*$ | 0.057$^*$ | 0.055$^*$ | 0.071$^*$ | −0.022$^*$ | 0.041$^*$ | 0.075$^*$ | −0.022$^*$ | 0.071$^*$ |
| Whole | 341388 | 0.037$^*$ | 0.050$^*$ | 0.037$^*$ | 0.073$^*$ | −0.038$^*$ | 0.058$^*$ | 0.062$^*$ | −0.026$^*$ | 0.075$^*$ |

* Significant ($p < 0.001$)

In the regression for $F_{cited}$, two explanatory variables had negative values of the standardized partial regression coefficient: *FG* (figures) and *PC* (priority claims). Regarding *FG*, the regression coefficient was negative in most of the sections. In other words, as with simple correlations, the number of times a patent is cited and the number of figures it has are negatively related to each other. However, in some sections, the regression coefficient of *FG* was not statistically significant ($p < 0.001$) or *FG* was not selected by the stepwise method (E: fixed constructions and F: mechanical engineering, etc.). Concerning *PC*, the regression coefficient was negative, although the simple correlation of *PC* with $F_{cited}$ exhibited a positive value. The regression coefficient of *PC* was not significant or *PC* was not selected by the stepwise method in half of the eight sections (B: performing operations; transporting, C: chemistry; metallurgy, D: textiles; paper, and E: fixed constructions).

Looking at the whole data of patents being studied, in contrast to the aforementioned results relating to the simple correlation with $F_{cited}$, $VC_{citing}$ (classifications associated with backward citations) had a higher value of the regression coefficient than had *PG* (pages) and *TB* (tables). The regression coefficient of $VC_{citing}$ was highest among the eight explanatory variables. This implies that, if confounding between factors is taken into account, the number of classifications assigned to cited patents, i.e., the diversity of backward citations, has a larger influence on the amount of forward citations than have other factors. However, absolute values of the standardized partial regression coefficient were low in general. Also, while the regression was statistically significant ($p$ < 0.001) not only for the whole data but also for each section, the regression models were low in the coefficient of determination $R'^2$ and did not fit the data well. In section D (textiles; paper), $R'^2$ was somewhat higher than in the other sections, but its value was not more than 0.10. The difficulty in the linear regression, in which citation frequency is used as the response variable, would be due to the distribution of its values, the majority of which are zero as shown in Fig. 2. Considering this problem, in the following analysis, I discuss the contribution of each factor in explaining and predicting citation frequency through logistic regression.

## Multiple logistic regression for citation frequency

First, including all explanatory variables except $PC_c$ in the model, I conducted a multiple logistic regression analysis that discriminated patents being cited at least once. The regression was performed for each section as well as for all patents published in 1998. As a summary of the results, Table 4 presents significance levels for the regression model and values of the partial regression coefficient of each variable. The regression model was significant both for the whole data of patents and for every individual section ($p$ < 0.001). Regarding the partial regression coefficient, all the variables were significant for the whole data ($p$ < 0.001). However, looking at each section individually, we find that the partial regression coefficients of *VC* (classifications), *PG* (pages), and *FG* (figures) were not significant in more than one section.

Table 4 Results of the multiple logistic regression analysis

| | Level of significance of the model | Partial regression coefficient | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | ***IV*** | ***VC*** | ***PG*** | ***FG*** | ***TB*** | ***CL*** | ***PC*** | ***VC*** $_{citing}$ |
| A | $p$ < 0.001 | 0.193$^{*}$ | 0.018 | −0.019 | −0.081$^{*}$ | 0.269$^{*}$ | 0.209$^{*}$ | −0.626$^{*}$ | 0.331$^{*}$ |
| B | $p$ < 0.001 | 0.204$^{*}$ | 0.320$^{*}$ | 0.298$^{*}$ | −0.066$^{*}$ | 0.196$^{*}$ | 0.184$^{*}$ | −0.268$^{*}$ | 0.254$^{*}$ |
| C | $p$ < 0.001 | 0.243$^{*}$ | 0.095 | 0.016 | −0.014 | 0.146$^{*}$ | 0.157$^{*}$ | −0.261$^{*}$ | 0.283$^{*}$ |
| D | $p$ < 0.001 | 0.429$^{*}$ | 0.287$^{*}$ | 0.221 | −0.243$^{*}$ | 0.246$^{*}$ | 0.116 | −0.587$^{*}$ | 0.208$^{*}$ |
| E | $p$ < 0.001 | 0.139$^{*}$ | 0.238$^{*}$ | 0.224$^{*}$ | −0.047 | 0.214$^{*}$ | 0.205$^{*}$ | −0.197 | 0.183$^{*}$ |
| F | $p$ < 0.001 | 0.239$^{*}$ | 0.364$^{*}$ | 0.324$^{*}$ | 0.081 | 0.223$^{*}$ | 0.108$^{*}$ | −0.394$^{*}$ | 0.289$^{*}$ |
| G | $p$ < 0.001 | 0.213$^{*}$ | 0.297$^{*}$ | 0.163$^{*}$ | 0.024 | 0.108$^{*}$ | 0.270$^{*}$ | −0.378$^{*}$ | 0.302$^{*}$ |
| H | $p$ < 0.001 | 0.241$^{*}$ | 0.368$^{*}$ | 0.224$^{*}$ | 0.009 | 0.142$^{*}$ | 0.225$^{*}$ | −0.404$^{*}$ | 0.290$^{*}$ |
| Whole | $p$ < 0.001 | 0.226$^{*}$ | 0.239$^{*}$ | 0.194$^{*}$ | −0.033$^{*}$ | 0.193$^{*}$ | 0.220$^{*}$ | −0.412$^{*}$ | 0.283$^{*}$ |

* Significant ($p$ < 0.001)

As in the case with the results of the linear regression mentioned previously, *FG* (figures) and *PC* (priority claims) had negative values of the partial regression coefficient. The latter had a negative value in every section (eight sections) while the former in over half of the sections (five sections). In addition, *PG* (pages) had a negative value only in section A (human necessities). Variables other than these three had positive values of the partial regression coefficient in all the sections.

Regarding each of the variables, I identified the section in which the partial regression coefficient had the highest absolute value, that is, the section in which the variable showed the highest influence on the citation frequency among the eight sections. The results were as follows: *TB* (tables), *PC* (priority claims), and $VC_{citing}$ (classifications associated with backward citations) became highest in section A (human necessities); *IV* (inventors) and *FG* (figures) did in section D (textiles; paper); *PG* (pages) did in section F (mechanical engineering, etc.); *CL* (claims) did in section G (physics); and *VC* (classifications) did in section H (electricity).

Table 5 shows the ratio of patents whose citation was correctly predicted by the above regression against all patents, that is to say, the rate of correct discrimination. The rate of correct discrimination was around 70%, both as a whole and in every section. The rate was comparatively high in section E (fixed constructions), at 73.9%, while it was comparatively low in section C (chemistry; metallurgy), at 64.3%. Nevertheless, no major differences were observed among the sections.

The majority of patents are comprised of those that are not cited at all, as indicated in Fig. 2. Accordingly, even if we would predict that all patents received no citations without subjecting them to the regression analysis, we could still obtain a high rate of correct discrimination. For example, in the case that 75% of all patents consist of those with the citation frequency of zero, predicting that “all patents receive no citations” without using the regression naturally results in a rate of correct discrimination reaching 75%. Therefore, the rate of correct discrimination does not sufficiently function as a measure for expressing the fitness and performance of models in this kind of regression. Hence, I evaluated regression models from the viewpoint of how precisely they could detect the minority of patents, namely those for which the citation frequency was equal to or beyond a certain threshold. Table 5 also shows the precision in detecting patents that had been cited once or more ($PR_1$). Along with $PR_1$, the following values are presented in Table 5: the number of patents $n$, the number of patents cited once or more $n_1$, and their ratio $n_1/n$. The ratio $n_1/n$ corresponds to the expected value of the precision obtained when patents are extracted at random. While the probability that the citation frequency of a randomly extracted patent is equal to or greater than 1 is not more than around 30% ($n_1/n$), predictions based on the regression model enable detection with a precision of around 50% ($PR_1$).

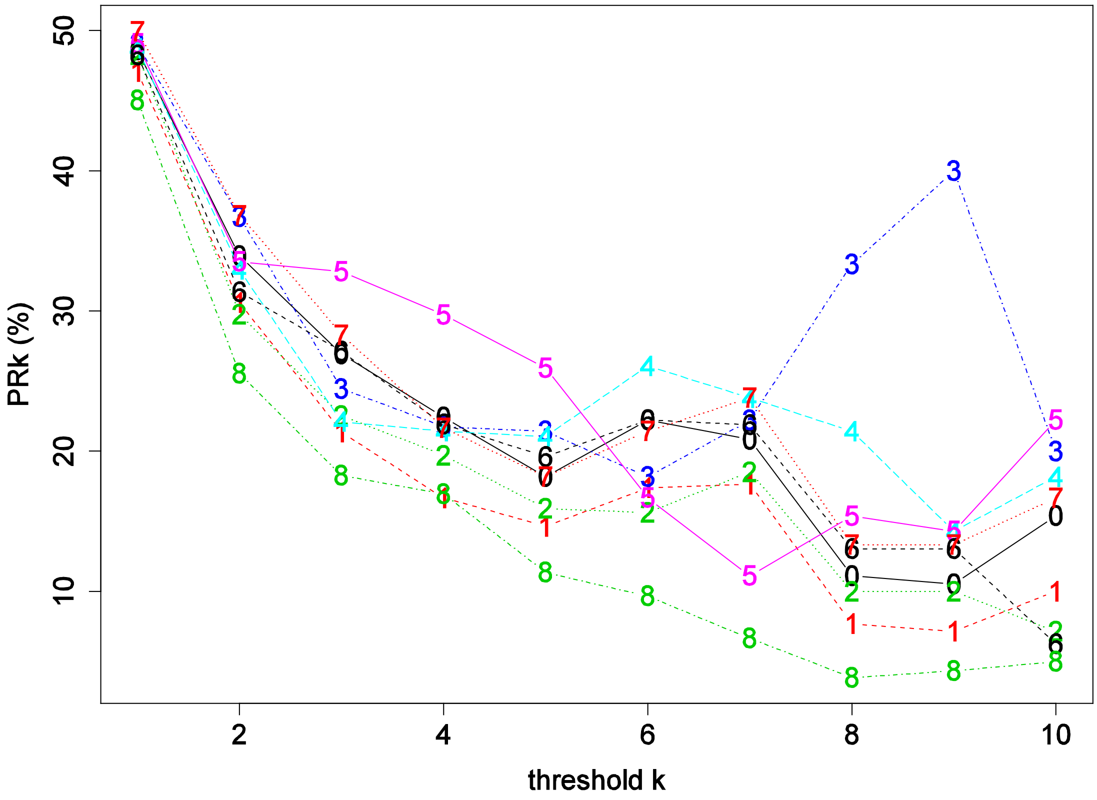
Table 5 Performance of the regression model

| | **Rate of correct discrimination (%)** | $PR_1$ **(%)** | $n$ | $n_1$ | $n_1/n$ **(%)** |
|---|---|---|---|---|---|
| A | 68.8 | 50.0 | 39474 | 12305 | 31.2 |
| B | 70.5 | 52.5 | 92483 | 27381 | 29.6 |
| C | 64.3 | 51.9 | 46881 | 16809 | 35.9 |
| D | 68.7 | 52.4 | 6255 | 1977 | 31.6 |
| E | 73.9 | 39.5 | 22564 | 5871 | 26.0 |
| F | 72.6 | 53.9 | 40437 | 11123 | 27.5 |
| G | 68.3 | 51.1 | 103919 | 32966 | 31.7 |
| H | 70.2 | 52.8 | 96679 | 28888 | 29.9 |
| Whole | 70.3 | 48.6 | 341388 | 101391 | 29.7 |

Next, conducting the regression analysis on the basis of each model in which one of the explanatory variables was excluded (in other words, the other seven variables were included), I discriminated patents whose citation frequency was equal to or beyond threshold $k$. The precision, $PR_k$, of those regression models was calculated and compared. On the basis of the comparison results, I now discuss the influence of excluding each variable on the regression. Transitions of precision $PR_k$ depending on the value of threshold $k$, which separates patents to be detected from the others, are illustrated for each regression model in Fig. 3. In addition, values of the precision when threshold $k$ is fixed at 1, i.e., $PR_1$, are illustrated for each section and each regression model in Fig. 4. In these figures, while symbol “0” refers to the model that employs all the eight variables, “1”–“8” refer to models that exclude each of the following, respectively: $IV$ (inventors), $VC$ (classifications), $PG$ (pages), $FG$ (figures), $TB$ (tables), $CL$ (claims), $PC$ (priority claims), and $VC_{citing}$ (classifications associated with backward citations).

In respect to transitions of precision $PR_k$ according to changes in threshold $k$, although there were exceptions such as model 3 (the model excluding $PG$), which had a high value of $PR_k$ at $k = 9$, we can observe as a general trend that the precision is lowered by the increase in the threshold value (Fig. 3). On the other hand, in respect to comparisons of the sections, what is noticeable is that values of precision $PR_1$ in section E (fixed constructions) were very low (Fig. 4). This would be affected by the fact that the proportion of patents to be detected, namely patents cited at least once, in section E is smaller than in the other sections (26.0%), as presented in Table 5.

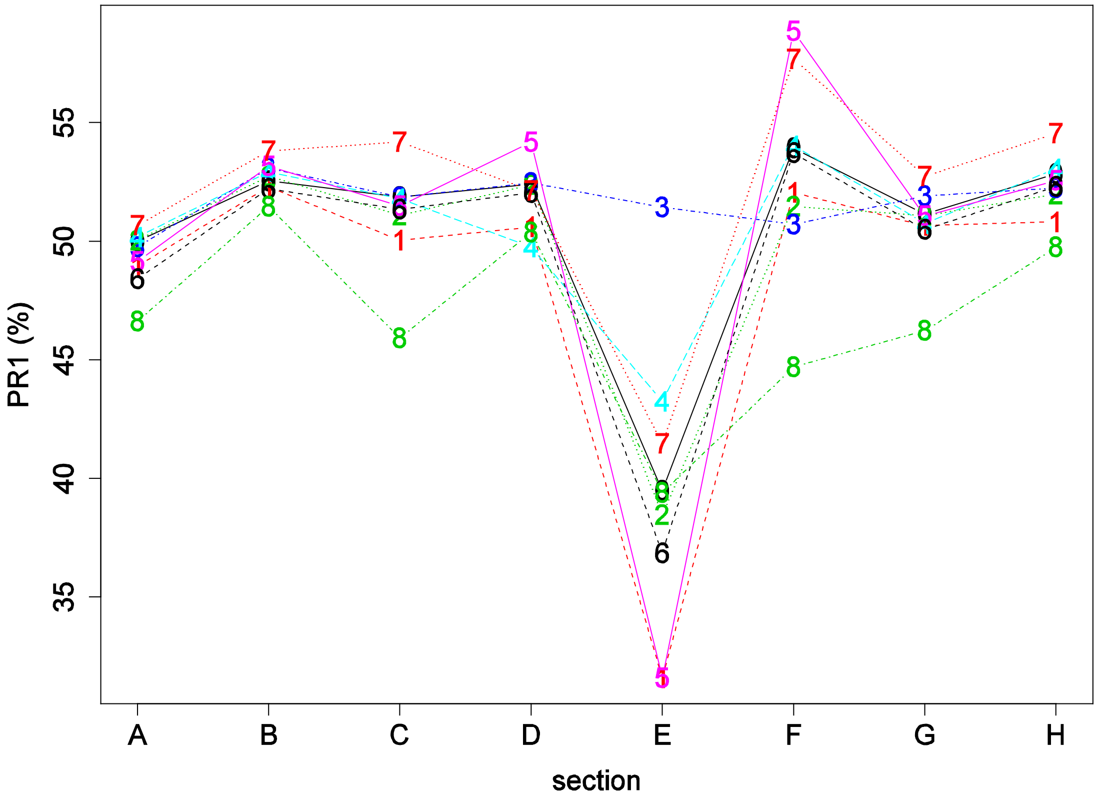
A comparison of the models that exclude each variable, as shown in Figs. 3 and 4, reveals that the order of the models' precision differs from threshold to threshold and from section to section at which they are compared. For instance, at around *k* = 4, the precision of model 5 (the model excluding *TB*) was highest, but at around *k* = 8, that of model 3 (the model excluding *PG*) was highest among all the models. However, it is common to all that among the nine models (model 0, which employs all variables, and models 1–8, which exclude each variable), model 8, which excludes $VC_{citing}$ (classifications associated with backward citations), had the lowest value or a value near it for the precision. With most of the threshold values and in most sections, the precision was at its lowest when $VC_{citing}$ was not incorporated into the regression model. Thus, we can confirm that the exclusion of $VC_{citing}$ tends to largely reduce the precision in the regression in which conditions of $F_{cited}$ are predicted, that is to say, the diversity of backward citations of a particular patent largely influences the number of its forward citations compared with other factors.



0: all, 1: excluding *IV*, 2: excluding *VC*, 3: excluding *PG*, 4: excluding *FG*,

5: excluding *TB*, 6: excluding *CL*, 7: excluding *PC*, 8: excluding $VC_{citing}$

**Fig. 3** Precision of the regression models for each threshold value



0: all, 1: excluding *IV*, 2: excluding *VC*, 3: excluding *PG*, 4: excluding *FG*,

5: excluding *TB*, 6: excluding *CL*, 7: excluding *PC*, 8: excluding $VC_{citing}$

**Fig. 4** Precision of the regression models for each section

## ZINB regression for citation frequency

Lastly, applying ZINB regression, I reconfirmed the correlation between $VC_{citing}$ (classifications associated with backward citations) and $F_{cited}$ (citation frequency). Values of the partial regression coefficient and standard error are presented for each variable in Table 6. The results showed that the partial regression coefficient of $VC_{citing}$ was statistically significant not only for the whole data ($p <$ 0.001), but also for all individual sections ($p < 0.001$ for sections except E, and $p < 0.05$ for section E).

As for the other variables, under the 5% significance level ($p < 0.05$), *FG* (figures) was not significant in section E, *CL* (claims) was not significant in section D, and *PC* (priority claims) was not significant in sections A and E. In addition, under the 0.1% significance level ($p < 0.001$), not only *PC* but also *TB* (tables) was not significant in multiple sections (three sections for *TB* while four sections for *PC*). As with in the linear and logistic analyses, negative values of the partial regression coefficient were observed for *FG*. On the other hand, the partial regression coefficient of *PC*, which was negative in the linear and logistic analyses, had positive values for all sections except A as well as for the whole data.

Table 6 Results of the ZINB regression analysis

| | Partial regression coefficient<br>(Standard error) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | ***IV*** | ***VC*** | ***PG*** | ***FG*** | ***TB*** | ***CL*** | ***PC*** | ***$VC_{citing}$*** |
| A | 0.257 | 0.210 | 0.452 | −0.189 | 0.141 | 0.169 | −0.065 | 0.142 |
| | (0.0327)$^{*}$ | (0.0384)$^{*}$ | (0.0392)$^{*}$ | (0.0172)$^{*}$ | (0.0259)$^{*}$ | (0.0252)$^{*}$ | (0.0608) | (0.0298)$^{*}$ |
| B | 0.270 | 0.336 | 0.675 | −0.265 | 0.270 | 0.183 | 0.251 | 0.367 |
| | (0.0203)$^{*}$ | (0.0247)$^{*}$ | (0.0257)$^{*}$ | (0.0125)$^{*}$ | (0.0194)$^{*}$ | (0.0152)$^{*}$ | (0.0376)$^{*}$ | (0.0164)$^{*}$ |
| C | 0.285 | 0.367 | 0.691 | −0.156 | 0.044 | 0.106 | 0.174 | 0.281 |
| | (0.0303)$^{*}$ | (0.0331)$^{*}$ | (0.0311)$^{*}$ | (0.0134)$^{*}$ | (0.0182)$^{+}$ | (0.0202)$^{*}$ | (0.0410)$^{*}$ | (0.0188)$^{*}$ |
| D | 0.398 | 0.888 | 0.835 | −0.382 | 0.236 | 0.005 | 0.338 | 0.353 |
| | (0.0876)$^{*}$ | (0.0929)$^{*}$ | (0.1196)$^{*}$ | (0.0420)$^{*}$ | (0.0672)$^{*}$ | (0.0723) | (0.1563)$^{+}$ | (0.0596)$^{*}$ |
| E | 0.269 | 0.251 | 0.309 | −0.048 | 0.155 | 0.158 | 0.058 | 0.108 |
| | (0.0447)$^{*}$ | (0.0539)$^{*}$ | (0.0731)$^{*}$ | (0.0346) | (0.0549)$^{+}$ | (0.0395)$^{*}$ | (0.0889) | (0.0440)$^{+}$ |
| F | 0.203 | 0.228 | 0.622 | −0.095 | 0.116 | 0.140 | 0.178 | 0.218 |
| | (0.0319)$^{*}$ | (0.0404)$^{*}$ | (0.0522)$^{*}$ | (0.0278)$^{*}$ | (0.0401)$^{+}$ | (0.0265)$^{*}$ | (0.0671)$^{+}$ | (0.0295)$^{*}$ |
| G | 0.216 | 0.269 | 0.679 | −0.384 | 0.345 | 0.269 | 0.263 | 0.420 |
| | (0.0193)$^{*}$ | (0.0239)$^{*}$ | (0.0218)$^{*}$ | (0.0127)$^{*}$ | (0.0214)$^{*}$ | (0.0146)$^{*}$ | (0.0358)$^{*}$ | (0.0167)$^{*}$ |
| H | 0.278 | 0.323 | 0.393 | −0.225 | 0.125 | 0.335 | 0.286 | 0.331 |
| | (0.0201)$^{*}$ | (0.0257)$^{*}$ | (0.0282)$^{*}$ | (0.0152)$^{*}$ | (0.0222)$^{*}$ | (0.0155)$^{*}$ | (0.0377)$^{*}$ | (0.0177)$^{*}$ |
| Whole | 0.245 | 0.242 | 0.622 | −0.261 | 0.200 | 0.231 | 0.213 | 0.357 |
| | (0.0106)$^{*}$ | (0.0136)$^{*}$ | (0.0130)$^{*}$ | (0.0064)$^{*}$ | (0.0102)$^{*}$ | (0.0081)$^{*}$ | (0.0205)$^{*}$ | (0.0094)$^{*}$ |

* Significant ($p < 0.001$)

+ Significant ($p < 0.05$)

# Conclusions

This study conducted multiple regression analyses using eight explanatory variables—the numbers of inventors, classifications, pages, figures, tables, claims, priority claims, and cited patents' classifications—for the following purposes: (1) to examine the hypothesis that the number of cited patents' classifications is positively related to the citation frequency under the condition that other factors are controlled and (2) to compare the influence of the number of cited patents' classifications with that of other factors based on the contribution of variables in the regression models. The main findings are as follows:

(1) The number of classifications of cited patents exhibits a statistically significant positive correlation with the citation frequency in linear, logistic, and ZINB models ($p < 0.001$).

(2) While in the observation of simple correlations, the numbers of pages and tables show the strongest relationship with the citation frequency, the multiple regression analyses demonstrate that the number of classifications of cited patents contributes more to the regression than do other factors, including the numbers of pages and tables, both for linear and logistic models.

The results of (2) imply that, if confounding between factors is taken into account, it is the diversity of classifications assigned to backward citations that has a larger influence on the number of forward citations.

In future studies, with the aim of clarifying the factors that affect the citation frequency of patents, I would like to conduct a more detailed analysis, such as a regression analysis on the basis of the measurement of the citation frequency in which positive citations are distinguished from negative ones.

### Acknowledgments

This work was partially supported by Grant-in-Aid for Scientific Research (C) 23500294 (2012) from the Ministry of Education, Culture, Sports, Science and Technology, Japan, and I would like to express my gratitude to the support. I also acknowledge the anonymous reviewers for their helpful comments.

## References

Bornmann, L., & Daniel, H. -D. (2007). Multiple publication on a single research study: Does it pay? The influence of number of research articles on total citation counts in biomedicine. *Journal of the American Society for Information Science and Technology, 58*(8), 1100–1107.

Bornmann, L., & Daniel, H. -D. (2008). What do citation counts measure?: A review of studies on citing behavior. *Journal of Documentation, 64*(1), 45–80.

Foltz, J., Barham, B., & Kim, K. (2000). Universities and agricultural biotechnology patent production. *Agribusiness, 16*(1), 82–95.

Fujii, A., Utiyama, M., Yamamoto, M., Utsuro, T., Ehara, T., Echizen-ya, H., & Shimohata, S. (2010). Overview of the patent translation task at the NTCIR-8 workshop. In: *Proceedings of NTCIR-8 workshop meeting* (pp. 371–376). Tokyo: National Institute of Informatics.

Glänzel, W. (2002). Co-authorship patterns and trends in the sciences (1980–1998): A bibliometric study with implications for database indexing and search strategies. *Library Trends, 50*(3), 461–473.

Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented inventions. *The Review of Economics and Statistics, 81*(3), 511–515.

Inuzuka, A. (2011). Factors facilitating technology reuse: An estimation from patent citation data. *Okayama Economic Review, 43*(3), 15–28.

Kostoff, R. N. (2007). The difference between highly and poorly cited medical articles in the journal Lancet. *Scientometrics, 72*(3), 513–520.

Lee, Y. -G., Lee, J. -D., Song, Y. -I., & Lee, S. -J. (2007). An in-depth empirical analysis of patent citation counts using zero-inflated count data model: The case of KIST. *Scientometrics*, *70*(1), 27–39.

Nanba, H., Fujii, A., Iwayama, M., & Hashimoto, T. (2008). Overview of the patent mining task at the NTCIR-7 workshop. In: *Proceedings of NTCIR-7 workshop meeting* (pp. 325–332). Tokyo: National Institute of Informatics.

Narin, F. (1995). Patents as indicators for the evaluation of industrial research output. *Scientometrics, 34*(3), 489–496.

Odagiri, H., Koga, T., & Nakamura, K. (2002). *R&D Boundaries of the Firm and the Intellectual Property System (Discussion Paper, 24)*. Tokyo: National Institute of Science and Technology Policy.

Peters, H. P. F., & van Raan, A. F. J. (1994). On determinants of citation scores: A case study in chemical engineering. *Journal of the American Society for Information Science, 45*(1), 39–49.

Sato, Y., & Iwayama, M. (2006). A study of patent document score based on citation analysis. *Information Processing Society of Japan SIG Technical Report, 2006*(59), 9–16.

Snizek, W. E., Oehler, K., & Mullins, N. C. (1991). Textual and nontextual characteristics of scientific papers: Neglected science indicators. *Scientometrics, 20*(1), 25–35.

Tang, L., & Shapira, P. (2012). Effects of international collaboration and knowledge moderation on China's nanotechnology research impacts. *Journal of Technology Management in China*, *7*(1), 94–110.

WIPO (World Intellectual Property Organization) (2010). *International Patent Classification (IPC)* [Web Page], Available: http://www.wipo.int/classifications/ipc/en/

Yoshikane, F., Suzuki, Y., & Tsuji, K. (2012). Analysis of the relationship between citation frequency of patents and diversity of their backward citations for Japanese patents. *Scientometrics, 92*(3), 721–733.