Unclassified **ENV/WKP(2015)2**

Organisation de Coopération et de Développement Économiques
Organisation for Economic Co-operation and Development **10-Feb-2015**

English - Or. English

ENVIRONMENT DIRECTORATE

ENV/WKP(2015)2
Unclassified

THE BENEFITS OF INTERNATIONAL CO-AUTHORSHIP IN SCIENTIFIC PAPERS: THE CASE OF WIND ENERGY TECHNOLOGIES - ENVIRONMENT WORKING PAPER No. 81

by Julie Poirier (1), Nick Johnstone (2), Ivan Haščič (3) and Jérôme Silva (3)

(1) Paris Graduate School of Economics, Statistics and Finance (ENSAE-ParisTech); Université Lyon 2
(2) OECD Directorate for Science, Technology and Innovation
(3) OECD Environment Directorate

Authorised for publication by Simon Upton, Director, Environment Directorate.

Keywords: climate change mitigation; scientific collaboration; innovation; knowledge spillovers

JEL classification: O3; O31; O38; Q4; Q42; Q48; Q55

OECD Environment working papers are available at www.oecd.org/env/workingpapers.htm

JT03370333

Complete document available on OLIS in its original format

English - Or. English

OECD ENVIRONMENT WORKING PAPERS

This series is designed to make available to a wider readership selected studies on environmental issues prepared for use within the OECD. Authorship is usually collective, but principal author(s) are named. The papers are generally available only in their original language - English or French - with a summary in the other language.

Comments on Working Papers are welcomed, and may be sent to:

Environment Directorate,
OECD, 2 rue André-Pascal, 75775 Paris Cedex 16, France
or by e-mail: env.contact@oecd.org

OECD Environment Working Papers are published on
www.oecd.org/env/workingpapers.htm

ABSTRACT

This paper presents an analysis of the effect of international co-authorship of scientific publications on patenting in wind energy technologies. It is found that the number of scientific publications co-authored by researchers in OECD countries has a positive and very significant impact on the number of wind energy innovations patented in OECD countries. However, non-OECD countries produce a greater number of patent filings when their researchers collaborate with OECD countries. This suggests that there exist knowledge spillovers between OECD and non-OECD countries that particularly benefit non-OECD countries. This empirical finding is important because it strengthens the case for international research cooperation between OECD and non-OECD countries in the area of climate mitigation.

***Keywords**:* climate change mitigation; scientific collaboration; innovation; knowledge spillovers

***JEL classification**:* O3; O31; O38; Q4; Q42; Q48; Q55

RESUME

On trouvera dans le présent document une analyse de l'incidence que le co-autorat international de publications scientifiques a sur le brevetage des technologies éoliennes. Il apparaît que le nombre de publications scientifiques rédigées conjointement par des chercheurs de la région OCDE a un impact positif et très significatif sur le nombre des innovations brevetées par les pays membres dans le domaine de l'énergie éolienne. Toutefois, on observe également que les pays non membres sont à l'origine d'un plus grand nombre de demandes de brevets lorsque les chercheurs de ces pays collaborent avec des homologues de pays de l'OCDE. Cela laisse penser qu'un transfert indirect de connaissances s'opère entre les pays membres et non membres de l'OCDE, principalement pour le bénéfice de ces derniers. Cette constatation empirique est importante car elle apporte un argument supplémentaire en faveur de la coopération entre chercheurs des pays membres et non membres de l'OCDE dans le domaine de l'atténuation du changement climatique.

Mots-clés : atténuation du changement climatique ; collaboration scientifique ; innovation ; diffusion des connaissances.

Classification JEL : O3; O31; O38; Q4; Q42; Q48; Q55

FOREWORD

This paper is a contribution to the OECD project on “Environmental Policy and Technological Innovation”. It has been authored by Julie Poirier (Paris Graduate School of Economics, Statistics and Finance ENSAE-ParisTech; Université Lyon 2), Nick Johnstone (OECD Directorate for Science, Technology and Innovation), Ivan Haščič and Jérôme Silva (OECD Environment Directorate). A draft of this paper was reviewed by the Working Party on Climate, Investment and Development (WPCID) and benefited from comments received. The authors are grateful to Jenny Calder and Elvira Berrueta-Imaz for editorial assistance.

EXECUTIVE SUMMARY

Encouraging international research is often seen as a potential source of major innovation breakthroughs, allowing researchers to access knowledge that is less readily available at home. Innovation in the development of “green technologies” can contribute to global challenges such as climate change. Since all countries can benefit from meeting such challenges, there is a particularly strong economic case for encouraging international research cooperation in the development of the relevant technologies.

Analysis of bibliometric, patent and other administrative data can help better understand the innovation system and identify the extent of collaboration in the innovation process. The evidence to-date shows that international scientific collaboration results in research with high impact and that the broader the collaboration, the higher the impact of the research. Small countries are generally more likely to engage in international collaboration than larger ones although this is not always the case and there are differences among disciplines. Indicators based on patents and scientific publications can be used to measure these phenomena.

In this paper the relationship between international research collaboration and patenting of inventions is examined in the field of wind energy technologies. To measure international research collaboration the analysis draws upon the SCOPUS database to construct a measure of international co-authorship of scientific publications, attributed to countries by the author's institutional affiliation at the time of publication. To measure patenting activity the paper draws on the PATSTAT database to construct a measure of patented inventions, attributed to countries by the residence of the inventor. The paper places a particular emphasis on examining the possible impacts of co-authorship between OECD and non-OECD researchers on the production of patents in OECD and non-OECD countries.

The principal research question is whether international co-authorship of scientific publications can influence patenting. Is scholarly work (i.e. publications) an input for patents, and how is this affected by international cooperation? The related hypotheses are tested separately for the sample of 29 OECD countries and the sample of 50 non-OECD countries during the period of 12 years (1998-2009).

It is found that the number of scientific publications co-authored by researchers in OECD countries has a positive and very significant impact on the number of wind energy innovations patented in OECD countries. In other words, **OECD countries innovate more when their scientists and researchers collaborate** with each other.

However, the key finding of this paper is that non-OECD countries produce a greater number of patent filings when their researchers collaborate with those in OECD countries. This suggests that there exist **knowledge spillovers between OECD and non-OECD countries that particularly benefit non-OECD countries**. This empirical finding is important because it strengthens the case for international research cooperation between OECD and non-OECD countries in the area of climate mitigation.

To the extent that all countries benefit from ‘pushing out the frontier’ in climate change mitigation innovation this is of mutual benefit. Policy initiatives such as the IEA’s Implementing Agreements, researcher mobility programmes, and non-discriminatory access to research grants should be considered.

TABLE OF CONTENTS

THE BENEFITS OF INTERNATIONAL CO-AUTHORSHIP IN SCIENTIFIC PAPERS: THE CASE OF WIND ENERGY TECHNOLOGIES

1. Introduction

Encouraging international research is often seen as a potential source of major innovation breakthroughs, allowing researchers to access knowledge that is less readily available at home. This implies developing networks of researchers across countries. What are the returns to innovation when different actors in different places create new knowledge? How can we measure the international transmission mechanisms of new knowledge and their impacts on innovation?

Investments, and thus innovation, in the development of "green technologies" can contribute to global challenges such as climate change. Since all countries can benefit from meeting such challenges, there is a particularly strong economic case for international research cooperation in relevant technologies. In this paper, the analysis is restricted to wind energy technologies.

There is a dual externality associated with environmental innovation. On the one hand, pollution is a negative externality (since elements of the assimilative capacity of the environment are public goods) while innovation is viewed as a positive externality (since elements of the information generated by innovation are public goods). Investment in technological innovation for climate change mitigation is likely to increase as many OECD countries implement binding national policies. However, reaching agreement on emission cuts at the international level would certainly provide a significant spur to innovation. At the Durban meeting of the Conference of the Parties to the United Nations Framework Convention on Climate Change parties agreed to collectively engage in a global legally binding agreement by 2015, which would enter into force in 2020. This could stimulate investment in clean technologies to reduce greenhouse gases emissions.

In this paper the relationship between international co-authorship of scientific publications and patents is examined. The paper places a particular emphasis on examining the possible impacts of co-authorship between OECD and non-OECD researchers on the production of patents in OECD and non-OECD countries. The following section presents a description of the data. This is followed by a detailed discussion of the modelling strategy adopted in this paper. Finally, the paper provides the empirical results and discusses their implications for policy.

2. Collaboration and the innovation process

The production of new knowledge is often a collective process involving a significant number of individuals and organizations which requires communication and coordination. Knowledge produced in such a complex but structured way may have public good aspects. Such interactions or networks may be usefully tracked as part of the innovation measurement framework. Networks can be a means for collective intelligence and policies that seek to influence the rate and orientation of innovation have to take networks into account. For instance, technology transfer between universities and industries implies two-way communication. The mobility of the highly skilled labour implies knowledge flows across disciplines, sectors and borders.

Analysis of bibliometric, patent and other administrative data can help reveal how these transnational networks are evolving. However, it should be emphasized that while science and innovation activities increasingly rely on dispersed networks of actors, they sometimes tend to cluster in certain places or around certain institutions (e.g. a leading university or a research laboratory of a multinational corporation). To analyse the changing landscape of science, technology and innovation is likely to require new units of analysis with different geographical scope.

A great deal of research has been devoted toward understanding how actors are linked in the innovation system and identifying how collaborative is the innovation process. Some papers in the network literature focus on knowledge flows (Choi 2012; Gao et al. 2011), other study the links between productivity and knowledge. For example, Jaffe (1989) developed a model of the regional knowledge production function, and Cho et al. (2010) were particularly interested in the relationship between productivity and co-authorship. Yet another stream of the literature analyses collaboration in patenting. For example, Jaffe et al. (1993) studied the geographic localization of knowledge spillovers using patent citations. Guan and Chen (2012) studied the links between patent collaboration and international knowledge flows. Franz (2010) compared the geographical locations of co-inventors of patents and of co-authors of scientific publications.

At the firm level, collaboration with foreign partners can play an important role in the innovation process by allowing firms to gain access to a broader pool of resources and knowledge at lower cost and risk. It can take a variety of forms and levels of interaction ranging from simple one-way information flows to highly interactive and formal arrangements. Collaboration rates vary widely across countries. In some countries, collaboration mainly involves national partners (e.g. Korea, China, Australia, Chile), but in most cases there is a greater balance between national and foreign partners. In some countries, firms are strongly oriented towards international collaboration (e.g. Luxembourg, the Slovak Republic, Finland and Switzerland).

Firm size is a strong determinant of foreign collaboration (Eurostat, CIS-2008[1]): large firms have a much higher propensity to collaborate internationally than SMEs (usually twice to three times as much), but in Australia, the United Kingdom and Israel the gap is narrower. In Korea, Brazil, China and Spain, which have relatively low international collaboration rates, there is almost no participation by SMEs. Among European firms, intra-European collaboration remains the predominant form of cross-country cooperation on innovation. In terms of collaboration outside Europe, European firms tend to partner mainly with US firms, although collaboration with firms in China and India is significant in Sweden, Finland and Belgium.

The evidence shows that international scientific collaboration results in research with high impact (as measured by citations) and that the broader the collaboration, the higher the impact of the research. Small countries are generally more likely to engage in international collaboration than larger ones although this is not always the case and there are differences among disciplines. Indicators based on patents and scientific publications can be used to measure these phenomena (Jaffe et al. 2005; Thompson et al. 2005).

3. Data

In this paper the relationship between published research and patented inventions is analysed. The analysis draws upon the SCOPUS database which contains data on research publications, and the PATSTAT database which includes data on patent documents from patent offices worldwide.

3.1 Some generalities on patents and the PATSTAT database

For the construction of our measure of invention data have been extracted from the EPO/OECD Worldwide Patent Statistical Database (usually referred to as PATSTAT). PATSTAT is a relational

[1] The Community innovation survey (CIS) collects information about product and process innovation, as well as organizational and marketing innovations.

database developed by the EPO from its master documentation database (DOCDB). It contains 20 separate tables including bibliographic data, citations and patent family links. It is designed to be used for statistical research, with data from over 100 countries and a total of 60 million patent applications. The PATSTAT database is an extensive and comprehensive database that answers the needs of researchers and policy-makers to combine different data sets for patent-related information.

The authors of this paper have developed search strategies to identify environmentally-relevant patent documents.[2] While the major OECD economies are generally the most active innovators in air and water pollution abatement and solid waste management, some smaller economies have developed specializations in this area. Work undertaken at the OECD indicates that predictability, flexibility and stringency of environmental policies are conducive to higher investment in innovation.

In the area of climate change mitigation this has been complemented with the efforts of patent examiners at the European Patent Office who have developed a tagging system for relevant technologies (the “Y02 tags”). Patents to address climate change challenges have been increasing (see e.g. Haščič et al. 2010). Patented innovation in these fields is concentrated in Germany, Japan and the United States, with different areas of specialization. For example Japanese patent applications are concentrated on innovation in energy, efficient buildings and lighting, as well as electric and hybrid vehicles. Innovation efforts in the United States focused particularly on renewable energy. Some countries have begun to invest considerable resources in advanced climate change mitigation technologies (e.g. solar photovoltaic energy, hydrogen and fuel cells, carbon capture and storage).

In PATSTAT it is possible to identify the country of residence of the inventors (and owners) of the patents. In many cases there will be multiple inventors for a given patent (co-invention). While inventors may live in different countries, domestic co-invention is much more common than international co-invention. This is consistent with the common finding in the literature that proximity is important for knowledge creation and technological progress (Choi 2012; Gao et al. 2011).

3.2 Some generalities on the SCOPUS database

The volume of scientific articles published is a key indicator as publication is the main means of disseminating and validating research results. Scopus is the world’s largest abstract and citation database of peer-reviewed literature. It contains about 20500 titles from 5000 publishers worldwide. In total it contains 49 million records, 78% with abstracts. Scopus excludes all documents for which the central purpose is not the presentation or discussion of scientific data, theory, methods, apparatus or experiments. Fields are determined by the classification of each journal.

Science and engineering include life science (clinical medicine, biomedical research and biology); physical science (chemistry, physics and Earth and space sciences); mathematics, social and behavioural sciences (social sciences, psychology, health sciences and professional fields). Finally engineering includes computer sciences and engineering and technology.

Variation across time and country in the incentives to publish raises a question of quality. Articles can be weighted by the frequency of citations. Citations attest to the productivity and influence of scientific literature (for example, there is a total of 35594 highly cited articles for 2006-2008, i.e. the top 1% of cited articles in the database for 2006-2008). In the Scopus database, articles were identified and distributed by country and type of collaboration.

[2] See http://www.oecd.org/env/consumption-innovation/indicator.htm ;
http://www.oecd.org/greengrowth/greengrowthindicators.htm

Collaboration is important for innovation at all stages of knowledge production. The increasing specialization of scientific disciplines and the increasing complexity of research encourage scientists to engage in collaborative research. There are new players on the research landscape. Moreover, there is increasing collaboration in sciences. Indeed, production of scientific knowledge is shifting from individuals to groups, from single to multiple institutions, and from national to international.

Researchers increasingly network across national and organizational borders. Publications are attributed to countries by the author's institutional affiliation at the time of publication. For this paper measures of international co-authorship have been developed based on affiliations of co-authors. Single authorship refers to scientific papers with a single author. Domestic co-authorship refers to scientific articles with two or more authors in the same country. International co-authorship refers to scientific articles with two or more authors from different countries. The classification is based on the number of addresses listed in each article.

International co-authorship is likely to be affected by language barriers and geographical factors. However, these obstacles have lessened as English has become the language most commonly used internationally by researchers. Physical distance between researchers is likely to have some correlation with the ratio of co-authorship, although the effect of information and communication technologies on knowledge flows has undoubtedly lessened its effect.

In most countries, research collaboration with foreign partners is at least as important as domestic cooperation. Actually, collaboration is likely to be undertaken to extend the scope of a project or to complement firm's competencies more than to save on costs. Therefore, policies that stimulate collaboration and network initiatives will have an impact on the entire spectrum of innovative firms.

Co-authorship of scientific articles provides a direct measure of collaboration in science. National and international co-authorship is far more prevalent than single authorship for all countries. International collaboration varies with country size. Small countries are generally more likely to engage in international collaboration than larger ones. However, when the number of scientific articles is taken into account, Germany, the United Kingdom and the United States attract the most international collaborations.

3.3 Construction of dataset

The focus of this paper is on the links between international co-authorship of scientific articles and patent applications. Because we are interested in innovation in climate change mitigation, the focus is on patents directed to environmental technologies. It would not be possible to define search strategies for the Scopus database for all environmental publications, and so the scope of the study was reduced to only the wind energy field.

The dependent variable is thus all the wind energy patent applications from 1996 to 2009 in both OECD and non-OECD countries. Extraction of wind power patent documents from the PATSTAT database was undertaken using the OECD Environment Directorate's extraction programme. Extraction from SCOPUS was made using keywords to identify only wind-energy publications. Given the narrowness of the field all types of articles published until the end of 2008 were used, and not only the top-cited articles as is often the case. We then matched both the publications and patents datasets.

In Scopus, authors are characterized by their university affiliation. Thus, we don't take into account authors' country of origin but rather authors' country of affiliation, and we classified the publications according to the country in which the university where the author is affiliated is located. However, some articles have multiple authors and some authors have multiple affiliations. So we had to define a strategy to allocate publications to countries.

In the literature, several matching methods have been proposed to affect a geographical location to authors and/or articles, in order to count the number of scientific publications produced by each country. Jaffe et al. (1993) used the following strategy to affect a unique location to each author and each publication. The principle is to assign each author/scientific paper to a country, based on pluralities of affiliations/authors. Ties are assigned randomly, except that ties between OECD and non-OECD countries are resolved in favour of non-OECD countries (Jaffe et al., 1993[10]; Thompson and Fox-Kean, 2005[12]). Another strategy is to consider each address as a separate observation (Furman et al., 2006[5]). According to this method, each affiliation (and as a result the geographical location associated with this affiliation) is considered as a separate observation.

Usually, when articles (or patents) have multiple authors (or inventors) from different countries, these articles (patents) are either partly attributed to each country mentioned (fractional count) or fully attributed to every relevant country (simple count), thus generating multiple counting at an aggregate level. In general, fractional counting procedures are used to compute counts by countries, but the alternative is sometimes preferable, as with indicators on international co-operation.

Because the research question addressed is the relative impact of OECD/and non-OECD research collaboration, the strategy adopted in this case differs from the two methods cited above. The methodology applied first consists of building for each publication all possible country pairs, according to each author's affiliation(s). Proceeding in this way allows for the assignment of each publication to every country where its authors are affiliated.

Finally, each country is classified as either OECD or non-OECD, taking into account the dates on which they deposited their instruments of ratification of the Convention on the Organization for Economic Co-operation and Development (OECD). We finally transform each country pair into a membership pair, according to countries' type of membership: OECD/OECD; non-OECD/non-OECD; and OECD/non-OECD. Notice that publications from a single country are not counted as co-authored publications but rather as domestic publications.

3.4 Descriptive statistics

After extraction of the data from the Scopus database, there are a total 14993 publications, with an average of a thousand articles published each year. Table 1 presents some descriptive statistics for this database.

Table 1. Descriptive statistics

Variable	Min	Max	Mean
Number of authors	1	47	2.90
Number of authors' affiliations	1	10	1.12
Number of authors' countries of affiliation	1	5	1.06
Number of countries per article	1	11	1.26

Overall, 71% of the authors in the database have an OECD affiliation and 29% have a non-OECD affiliation. Publications can be single-authored or co-authored, and different types of co-authorship are possible, including the following:

a) *Domestic publications* that include both single and nationally co-authored publications.

b) *OECD/non-OECD co-authored publications* that include publications with either at least one author from an OECD country and one author from a non-OECD country, or a single author which has several scientific affiliations in at least one non-OECD country and one OECD country.

c) *Non-OECD/non-OECD co-authored publications* that include publications with either at least two authors, each of them being from a non-OECD country, or only a single author who has scientific affiliations in several non-OECD countries.

d) *OECD/OECD co-authored publications* that include publications with either at least two authors, each of them being from an OECD country, or only a single author which has scientific affiliations in several OECD countries.

Each publication can be of several types of co-authorship, since it has several authors with different countries of affiliation. Numbers of publications' types are summarized in Table 2.

Table 2. Number of publications according to types of co-authorship

Variable	Description	Nb. of publications
PUBL_DOMESTIC	Number of publications with at least a domestic pair of authors	12251
PUBL_INTER_OECDOECD	Number of publications with at least an OECD pair of authors	3744
PUBL_INTER_OECDNON	Number of publications with at least an OECD/non-OECD pair of authors	2158
PUBL_INTER_NONNON	Number of publications with at least a non-OECD pair of authors	214
TOTAL_PUBLI	Total number of publications types	18367

After extraction of our patent data from the PATSTAT database, there are 2585 patents related to wind energy technologies in the database. On average, each country files 8 patent applications per year.

4. Estimation models

The principal hypothesis is that international co-authorship of scientific publications can influence patenting. Is scholarly work (i.e. publications) an input for patents, and how is this affected by international cooperation? Do authors in country i benefit more from co-authorship than from single authorship? This hypothesis is tested separately for OECD countries and non-OECD countries separately. Consequently two types of models are estimated – for patents registered in OECD countries and for patents registered in non-OECD countries.

In the case of non-OECD countries, innovation in wind energy is specified as follows:

$$PAT\_WIND_{it} = f(PUBL\_DOMESTIC_{it}; PUBL\_INTER\_OECDNON_{it}; PUBL\_INTER\_NONNON_{it}; \\ PAT\_TOTAL_{it}; \omega_t; \varepsilon_{it}) \quad (1)$$

where i indexes country and t indexes year.

The dependent variable (PAT_WIND) represents the number of patent applications in the wind energy class, classified by inventor country[3] and priority year[4]. The variable represents the number of unique simple patent families worldwide because only claimed priorities and singular applications are counted, hence avoiding double-counting of inventions[5].

As a control, a variable reflecting the propensity to invent and patent technologies in general (PAT_TOTAL) is included as an explanatory variable. It is constructed in a manner analogous to the dependent variable (a count of patent families by inventor country and priority year) with the difference being that all types of technologies (not only wind energy) are covered[6].

The number of domestic scientific publications (PUBL_DOMESTIC) is included as an explanatory variable. Assuming that a large number of domestic publications means a dynamic national R&D sector, the sign is expected to be positive. The data source is the Scopus database. The data covers the field of wind energy-related publications.

The number of OECD/non-OECD co-authored publications (PUBL_INTER_OECDNON) is included as an explanatory variable. Assuming that collaboration with at least one OECD country can boost innovation in non-OECD countries, the sign is expected to be positive. The data source is the Scopus database. The data covers the field of wind energy-related publications.

The number of non-OECD/non-OECD co-authored publications (PUBL_INTER_NONNON) is included as an explanatory variable. As for the previous variables, the data covers the field of wind energy-related publications.

Year fixed effects ω_t account for omitted year-variant effects that influence the dependent variable in a country-invariant manner. Most notably, this would capture the effect of conjectural or specific events that could influence the patenting activity. All the residual variation is captured by the error term ε_{it}.

In the case of OECD countries, innovation in wind energy is specified as follows:

$$PAT\_WIND_{it} = f(PUBL\_DOMESTIC_{it}; PUBL\_INTER\_OECDNON_{it}; PUBL\_INTER\_OECDOECD_{it}; PAT\_TOTAL_{it}; \omega_t; \varepsilon_{it}) \qquad (2)$$

where i indexes country and t indexes year.

The explanatory variables are the same as in the equation for non-OECD countries, except that the PUBL_INTER_NONNON variable is replaced by the PUBL_INTER_OECDOECD variable. The number of OECD/OECD co-authored publications (PUBL_INTER_OECDOECD) is thus included as an explanatory variable.

[3] The inventor country is the country of the residence of the inventor, which is frequently used to count patents in order to measure inventive performance (OECD, 2008).

[4] The priority year is the first year of filing of a patent application, anywhere in the world (normally in the applicant's domestic patent office) to protect an invention (OECD, 2008).

[5] Using data on patent families, the following types of documents are distinguished: singular is patent applied for at a single office, with no subsequent _lings elsewhere (i.e. patent family size= 1); claimed priority is patent for which an application is filed at an additional office to that of the priority office (these are inventions that have been applied for protection in multiple countries); and finally, duplicate is the additional application (OECD, 2009).

[6] This is achieved by extracting data on all patent applications with any IPC code assigned.

It is important to take into account the effect of publications in past years. For this reason, lagged variables are included in the innovations models. However, in order to avoid collinearity and endogeneity problems, three separate models are run for OECD and non-OECD countries, the first one including only contemporaneous (t) variables, the second one including only one-year lagged ($t-1$) variables and the third one including only two-year lagged ($t-2$) variables.

In all cases, the dependent variables represent the number of patent applications, that is to say patent counts. As discussed above, count data models, such as the Poisson and negative binomial, have been suggested for estimating the number of occurrences of an event, or event counts. Since there are a large number of zeros (73.4% in the non-OECD sample and 23.84% in the OECD sample), the zero-inflated variant of the negative binomial model is estimated in which the count process and the binary process are modelled separately.

Moreover, a number of other count data models, including some negative binomial models estimated using the Generalized Estimating Equations method are estimated. Tests are conducted to determine which model performs best. Maximum likelihood method is used to estimate the model parameters.

5. Results

The results of the estimation for the OECD sample are presented in Tables 3-5. The total sample (345 observations) is an unbalanced panel with 29 countries and 12 years (1998-2009). A series of tests were conducted to determine which model performs best.

Based on the link test, the negative binomial model performs better than the Poisson model (the dispersion parameter is significantly different from zero). Moreover, both the Akaike and the Bayesian information criterion (AIC, BIC) provide more robust evidence for the negative binomial model with year fixed effects and for the zero-inflated negative binomial model with year fixed effects. The values of the log-likelihood test go in the same direction. Based on these tests, the preferred models are the negative binomial model with year fixed effects and the zero-inflated negative binomial model with year fixed effects. That is why only those results are discussed in detail below.

The number of publications co-authored by OECD authors has a positive and very significant impact on the number of wind energy innovations patented in OECD countries (see Table 3). Patenting overall is also positive and very significant. The effect of domestic publications is statistically insignificant. Moreover, the number of publications co-authored by both OECD and non-OECD authors does not have any impact on wind energy patenting in OECD countries.

Sometimes when analysing a response variable that is a count variable, the number of zeros may seem excessive. In our OECD sample, the response variable (the number of wind energy patents filed in OECD countries) is a count, and the number of zeros in the OECD sample (23.8%) is high. It is important, therefore, to consider separately the processes that could lead to a response variable value of zero. A research group may have tried to generate patents, but failed to file them. Another group may have not tried to create any patent at all and, not surprisingly, filed zero patents. The first group could have filed one or more patents, but did not do so. The second group was certain to file zero patents. Thus, the number of zeros may be inflated and the number of groups filing zero patents cannot be explained in the same manner as the groups that filed more than zero patents.

A standard negative binomial model would not distinguish between the two processes causing an excessive number of zeroes, but a zero-inflated model allows for and accommodates this complication. When analysing a dataset with an excessive number of outcome zeros and two possible processes that arrive at a zero outcome, a zero-inflated model should be considered. Looking at the zero-inflated negative binomial model with year fixed effects, the number of zeros is predicted with both patenting overall and total number of patents. These two variables are statistically significant and have a negative impact on the number of zeros. Finally, year fixed effects are all statistically significant. When estimating the one-year lagged model (i.e. the model with explanatory variables taken in t-1, the results are similar (see Table 4).

When estimating the two-year lagged model, different results are obtained (see Table 5). The number of publications co-authored by OECD authors and patenting overall still have positive and very significant impacts on the number of patented wind energy innovations. However, the number of domestic publications becomes statistically significant with a positive impact on patenting in wind energy technologies. Moreover, the number of OECD/non-OECD co-authored publications has a negative influence on wind energy patents which are filed in OECD countries. Finally, the two explanatory variables of the inflation model are significantly negative again.

Table 3. Model estimates on the OECD sample (explanatory variables in t)

Dependent: PAT_WIND_it	**NegBin**	**NegBin with year fixed effects**	**NegBin (GEE method)**	**NegBin with year fixed effects (GEE method)**	**Poisson**	**Zero-Inflated NegBin**	**Zero-Inflated NegBin with year fixed effects**
Intercept	1.2321*** (0.1077)	-0.8812*** (0.3319)	1.2321*** (0.2172)	-0.8812*** (0.3328)	2.1470*** (0.0183)	1.6587*** (0.1058)	-0.3135 (0.2962)
Nbpubldomestic	-0.0005 (0.0050)	0.0006 (0.0047)	-0.0005 (0.0064)	0.0006 (0.0060)	-0.0056*** (0.0004)	-0.0002 (0.0042)	0.0010 (0.0039)
Nbpublinteroecdnon	-0.0185 (0.0231)	0.0032 (0.0214)	-0.0185 (0.0149)	0.0032 (0.0202)	-0.0266*** (0.0015)	-0.0107 (0.0192)	0.0091 (0.0180)
Nbpublinteroecdoecd	0.0769*** (0.0118)	0.0861*** (0.0113)	0.0769 (0.0116)	0.0861*** (0.0148)	0.0555*** (0.0012)	0.0599*** (0.0098)	0.0675*** (0.0093)
TotalPatents	0.0154*** (0.0030)	0.0122*** (0.0028)	0.0154*** (0.0056)	0.0122** (0.0049)	0.0082*** (0.0002)	0.0122*** (0.0025)	0.0093*** (0.0024)
Inflation model (logistic)							
Nbtotalpubli	-	-	-	-	-	-0.0714** (0.0320)	-0.1032*** (0.0375)
TotalPatents	-	-	-	-	-	-1.2678*** (0.3447)	-1.2919*** (0.3602)
Dispersion or Alpha	1.6032 (0.1397)	1.3977 (0.1251)	-	-	-	1.0602*** (0.1039)	0.9177*** (0.0898)
Year fixed effects	No	Yes	No	Yes	No	No	Yes
Nb of observations / Nb of zeros (when relevant)	345	345	345	345	345	345 / 82	345 / 82
Log Likelihood	-1135	-1114	-	-	-4756	-1096	-1073
AIC or QIC	2282	2262	-42285	-45689	9521	2210	2186
BIC	2305	2328	-	-	9540	2245	2262

*Significance levels: *=10%; **=5%; ***=1%*

Table 4. Model estimates on the OECD sample (explanatory variables in t-1)

Dependent: PAT_WIND_it	**NegBin**	**NegBin with year fixed effects**	**NegBin (GEE method)**	**NegBin with year fixed effects (GEE method)**	**Poisson**	**Zero-Inflated NegBin**	**Zero-Inflated NegBin with year fixed effects**
Intercept	1.2969*** (0.1050)	0.0414 (0.3157)	1.2969*** (0.2196)	0.0414 (0.3653)	2.1400*** (0.0177)	1.7120*** (0.1003)	0.5212* (0.2812)
Nbpubldomestic1	0.0025 (0.0055)	0.0026 (0.0054)	0.0025 (0.0112)	0.0026 (0.0099)	-0.0053*** (0.0006)	0.0018 (0.0046)	0.0024 (0.0044)
Nbpublinteroecdnon1	-0.0551* (0.0297)	-0.0173 (0.0312)	-0.0551*** (0.0178)	-0.0173 (0.0207)	-0.0417*** (0.0014)	-0.0456* (0.0241)	-0.0089 (0.0257)
Nbpublinteroecdoecd1	0.0876*** (0.0125)	0.0944*** (0.0131)	0.0876*** (0.0176)	0.0944*** (0.0192)	0.0681*** (0.0013)	0.0708*** (0.0100)	0.0739*** (0.0105)
TotalPatents1	0.0160*** (0.0030)	0.0130*** (0.0029)	0.0160** (0.0075)	0.0130** (0.0064)	0.0067*** (0.0003)	0.0127*** (0.0024)	0.0098*** (0.0024)
Inflation model (logistic)							
Nbtotalpubli1	-	-	-	-	-	-0.1198** (0.0482)	-0.1390*** (0.0515)
TotalPatents1	-	-	-	-	-	-1.2271*** (0.3385)	-1.2359*** (0.3381)
Dispersion or Alpha	1.5874 (0.1390)	1.4679 (0.1304)	-	-	-	1.0267*** (0.1000)	0.9413*** (0.0920)
Year fixed effects	No	Yes	No	Yes	No	No	Yes
Nb of observations / Nb of zeros (when relevant)	345	345	345	345	345	345 / 82	345 / 82
Log Likelihood	-1134	-1122	-	-	-4254	-1092	-1078
AIC or QIC	2280	2278	-42878	-43388	8519	2201	2196
BIC	2304	2344	-	-	8538	2236	2273

*Significance levels: *=10%; **=5%; ***=1%*

Table 5. Model estimates on the OECD sample (explanatory variables in t-2)

Dependent: PAT_WIND_it	**NegBin**	**NegBin with year fixed effects**	**NegBin (GEE method)**	**NegBin with year fixed effects (GEE method)**	**Poisson**	**Zero-Inflated NegBin**	**Zero-Inflated NegBin with year fixed effects**
Intercept	1.4306*** (0.1060)	-0.4851 (0.3118)	1.4306*** (0.2235)	-0.4851 (0.3321)	2.2946*** (0.0166)	1.7788*** (0.1018)	-0.0956 (0.2758)
Nbpubldomestic2	0.0075 (0.0068)	0.0137** (0.0064)	0.0075 (0.0155)	0.0137 (0.0149)	-0.0087*** (0.0007)	0.0049 (0.0056)	0.0116** (0.0053)
Nbpublinteroecdnon2	-0.1042*** (0.0385)	-0.1007** (0.0400)	-0.1042*** (0.0330)	-0.1007*** (0.0311)	-0.0321*** (0.0014)	-0.0876*** (0.0313)	-0.0820** (0.0328)
Nbpublinteroecdoecd2	0.0870*** (0.0153)	0.0927*** (0.0145)	0.0870*** (0.0171)	0.0927*** (0.0201)	0.0612*** (0.0015)	0.0741*** (0.0125)	0.0778*** (0.0118)
TotalPatents2	0.0182*** (0.0033)	0.0146*** (0.0031)	0.0182* (0.0093)	0.0146* (0.0088)	0.0107*** (0.0004)	0.0153*** (0.0028)	0.0112*** (0.0026)
Inflation model (logistic)							
Nbtotalpubli2	-	-	-	-	-	-0.2934** (0.1237)	-0.3146*** (0.1187)
TotalPatents2	-	-	-	-	-	-1.2248*** (0.3490)	-1.2268*** (0.3442)
Dispersion or Alpha	1.7304 (0.1483)	1.5268 (0.1346)	-	-	-	1.2030*** (0.1127)	1.0215*** (0.0978)
Year fixed effects	No	Yes	No	Yes	No	No	Yes
Nb of observations / Nb of zeros (when relevant)	345	345	345	345	345	345 / 82	345 / 82
Log Likelihood	-1147	-1128	-	-	-4940	-1109	-1085
AIC or QIC	2305	2290	-41762	-40022	9890	2235	2211
BIC	2329	2355	-	-	9909	2270	2288

*Significance levels: *=10%; **=5%; ***=1%*

The results of the estimation for the non-OECD sample are presented in Tables 6-8. The total panel (591 observations) is an unbalanced sample with 50 countries and 12 years (1998-2009). A series of tests were conducted to determine which model performs best. Based on the link test, the negative binomial model performs better than the Poisson model (the dispersion parameter is significantly different from zero). Moreover, both the Akaike and the Bayesian information criterion (AIC, BIC) provide more robust evidence for the negative binomial model with year fixed effects and for the zero-inflated negative binomial model with year fixed effects. The values of the log-likelihood test go in the same direction. Based on these tests, the preferred models are the negative binomial model with year fixed effects and the zero-inflated negative binomial model with year fixed effects. That is why only those results are discussed in detail.

The number of publications co-authored by both OECD and non-OECD authors has a positive and significant impact on the number of wind energy innovations patented in non-OECD countries (see Table 6). Patenting overall is also positive and very significant. Moreover, the number of publications co-authored by non-OECD authors and the number of domestic publications do not have any impact on wind energy patenting in non-OECD countries.

In the non-OECD sample, the response variable (the number of filed wind energy patents in non-OECD countries) is a count, and the number of zeros in this sample (73.4%) seem excessive. A zero-inflated model is implemented. Looking at the zero-inflated negative binomial model with year fixed effects, the number of zeros is predicted with both patenting overall and total number of patents. These two variables are statistically significant and have a negative impact on the number of zeros. Finally, year fixed effects are all statistically significant.

When estimating the one-year lagged model, results are similar for the negative binomial model with year fixed effects. However, when estimating the zero-inflated negative binomial model with year fixed effects, the total number of patents is still significantly positive but OECD/non-OECD co-authored publications is not significant anymore (see Table 7). The number of domestic publications becomes significant with a negative impact on wind energy patents filed in non-OECD countries. Finally, the number of publications co-authored by non-OECD authors is not significant.

When estimating the two-year lagged model, conclusions regarding the negative binomial model are still the same (see Table 8). With the zero-inflated negative binomial model, the number of domestic publications is statistically significant with a negative impact on patenting in wind energy technologies in non-OECD countries. The number of publications co-authored by both OECD and non-OECD authors and patenting overall still have positive and very significant impacts on the number of wind energy innovations patented in non-OECD countries. Hence, there could be some knowledge spillovers from OECD to non-OECD countries. Moreover, the number of non-OECD co-authored publications does not have any impact on wind energy patents which are _led in non-OECD countries. Finally, the two explanatory variables of the inflation model are significant and negative.

Thus, collaboration with OECD countries does play a positive role on the number of patents filed in non-OECD countries. The reverse is not true, that is to say collaboration with non-OECD countries does not seem to positively impact the number of patents filed in OECD countries, except when OECD/non-OECD scientific collaborations have taken place two years before the patents' filing date.

Table 6. Model estimates on the non-OECD sample (explanatory variables in t)

Dependent variable: PAT_WIND_it	**NegBin**	**NegBin with year fixed effects**	**NegBin (GEE method)**	**NegBin with year fixed effects (GEE method)**	**Poisson**	**Zero-Inflated NegBin**	**Zero-Inflated NegBin with year fixed effects**
Intercept	-0.9053*** (0.1337)	-0.5533 (0.3593)	-0.9053*** (0.1982)	-0.5533 (0.4202)	-0.2547*** (0.0469)	0.6604*** (0.1386)	0.9186** (0.3758)
Nbpubldomestic	0.0028 (0.0121)	-0.0091 (0.0082)	0.0028 (0.0030)	-0.0091*** (0.0028)	-0.0078*** (0.0010)	-0.0019 (0.0030)	-0.0013 (0.0026)
Nbpublinteroecdnon	0.1497*** (0.0437)	0.1077*** (0.0367)	0.1497*** (0.0167)	0.1077*** (0.0188)	0.0624*** (0.0053)	0.0507** (0.0200)	0.0387** (0.0178)
Nbpublinternonnon	-0.0294 (0.0960)	-0.0218 (0.0865)	-0.0294 (0.0573)	-0.0218 (0.0477)	0.0810*** (0.0195)	-0.0747 (0.0582)	-0.0941 (0.0624)
TotalPatents	0.0506*** (0.0110)	0.0534*** (0.0100)	0.0506*** (0.0059)	0.0534*** (0.0061)	0.0099*** (0.0008)	0.0115*** (0.0036)	0.0125*** (0.0038)
Inflation model (logistic)							
Nbtotalpubli	-	-	-	-	-	-0.1432** (0.0556)	-0.0874* (0.0447)
TotalPatents	-	-	-	-	-	-1.6168*** (0.5843)	-2.4068** (1.0284)
Dispersion / Alpha	4.5346 (0.5301)	3.7037 (0.4534)	-	-	-	1.3135*** (0.2268)	1.1021*** (0.2033)
Year fixed effects	No	Yes	No	Yes	No	No	Yes
Nb of observations / Nb of zeros	591	591	591	591	591	591 / 434	591 / 434
Log Likelihood	-625	-606	-	-	-11771	-583	-568
AIC / QIC	1261	1246	-557	-646	2364	1185	1175
BIC	1288	1320	-	-	2385	1224	1263

*Significance levels: *=10%; **=5%; ***=1%*

Table 7. Model estimates on the non-OECD sample (explanatory variables in t-1)

Dependent variable: PAT_WIND_it	**NegBin**	**NegBin with year fixed effects**	**NegBin (GEE method)**	**NegBin with year fixed effects (GEE method)**	**Poisson**	**Zero-Inflated NegBin**	**Zero-Inflated NegBin with year fixed effects**
Intercept	-0.8766*** (0.1320)	-0.5492 (0.3498)	-0.8766*** (0.1940)	-0.5492 (0.4678)	-0.2761*** (0.0473)	0.6753*** (0.1348)	0.8732*** (0.3229)
Nbpubldomestic1	0.0207 (0.0158)	0.0047 (0.0152)	0.0207* (0.0107)	0.0047 (0.0128)	-0.0220*** (0.0015)	-0.0113 (0.0072)	-0.0139** (0.0062)
Nbpublinteroecdnon1	0.0798* (0.0435)	0.0706* (0.0394)	0.0798*** (0.0193)	0.0706*** (0.0166)	0.0719*** (0.0054)	0.0384** (0.0193)	0.0299 (0.0185)
Nbpublinternonnon1	0.2819 (0.1807)	0.1007 (0.1620)	0.2819 (0.2079)	0.1007 (0.1528)	0.1084*** (0.0388)	0.0193 (0.1128)	-0.0794 (0.1082)
TotalPatents1	0.0500*** (0.0108)	0.0491*** (0.0101)	0.0500*** (0.0122)	0.0491*** (0.0134)	0.0205*** (0.0011)	0.0180*** (0.0049)	0.0204*** (0.0049)
Inflation model (logistic)							
Nbtotalpubli1	-	-	-	-	-	-0.1397** (0.0618)	-0.1026* (0.0576)
TotalPatents1	-	-	-	-	-	-2.0346*** (0.6357)	-2.5824*** (0.9148)
Dispersion / Alpha	4.5855 (0.5365)	3.9131 (0.4768)	-	-	-	1.3369*** (0.2340)	1.0989*** (0.1985)
Year fixed effects	No	Yes	No	Yes	No	No	Yes
Nb of observations / Nb of zeros	591	591	591	591	591	591 / 434	591 / 434
Log Likelihood	-626	-612	-	-	-1149	-587	-570
AIC / QIC	1265	1258	-604	-623	2308	1192	1179
BIC	1291	1332	-	-	2330	1231	1267

*Significance levels: *=10%; **=5%; ***=1%*

Table 8. Model estimates on the non-OECD sample (explanatory variables in t-2)

Dependent variable: PAT_WIND_it	**NegBin**	**NegBin with year fixed effects**	**NegBin (GEE method)**	**NegBin with year fixed effects (GEE method)**	**Poisson**	**Zero-Inflated NegBin**	**Zero-Inflated NegBin with year fixed effects**
Intercept	-0.8862***	-1.0499***	-0.8862***	-1.0499***	-0.3007***	0.6140***	0.5675*
	(0.1241)	(0.3780)	(0.1919)	(0.4065)	(0.0477)	(0.1339)	(0.3424)
Nbpubldomestic2	0.0230	0.0148	0.0230	0.0148	-0.0323***	-0.0178*	-0.0202**
	(0.0183)	(0.0188)	(0.0304)	(0.0297)	(0.0029)	(0.0100)	(0.0087)
Nbpublinteroecdnon2	0.1071**	0.0922**	0.1071***	0.0922***	0.0943***	0.0651***	0.0575***
	(0.0483)	(0.0440)	(0.0238)	(0.0209)	(0.0043)	(0.0195)	(0.0183)
Nbpublinternonnon2	0.2299	0.2151	0.2299	0.2151	0.0132	-0.0368	-0.0611
	(0.2084)	(0.2118)	(0.1441)	(0.1624)	(0.0426)	(0.1353)	(0.1374)
TotalPatents2	0.0543***	0.0503***	0.0543***	0.0503**	0.0265***	0.0224***	0.0221***
	(0.0103)	(0.0099)	(0.0194)	(0.0196)	(0.0013)	(0.0050)	(0.0048)
Inflation model (logistic)							
Nbtotalpubli2	-	-	-	-	-	-0.1858***	-0.1573**
						(0.0680)	(0.0632)
TotalPatents2	-	-	-	-	-	-2.0317***	-2.3103***
						(0.6078)	(0.6998)
Dispersion / Alpha	4.1599	3.5343	-	-	-	1.1550***	0.9456***
	(0.5049)	(0.4433)				(0.2117)	(0.1797)
Year fixed effects	No	Yes	No	Yes	No	No	Yes
Nb of observations / Nb of zeros	591	591	591	591	591	591 / 434	591 / 434
Log Likelihood	-619	-604	-	-	-1086	-577	-562
AIC / QIC	1251	1241	-499	-596	2182	1172	1164
BIC	1277	1316	-	-	2284	1212	1251

*Significance levels: *=10%; **=5%; ***=1%*

6. Conclusions

This paper reports on an analysis of the effect of international co-authorship of scientific publications on patenting in wind-energy technologies. More precisely, the effect of cooperation between researchers from OECD and non-OECD countries on patent counts is assessed. The PATSTAT and SCOPUS databases have been used in order to have data on both scientific publications and patent applications. The sample has been restricted to wind energy.

It is found that non-OECD countries produce a greater number of patent filings when their scientists and researchers collaborate with OECD countries. This suggests that there exist knowledge spillovers between OECD and non-OECD countries that particularly benefit non-OECD countries. This is important because it strengthens the case for international research cooperation between OECD and non-OECD countries in the area of climate mitigation. To the extent that all countries benefit from ‘pushing out the frontier’ in climate mitigation innovation this is of mutual benefit. Policy initiatives such as the IEA’s Implementing Agreements, researcher mobility programmes, and non-discriminatory access to research grant programmes should be considered.

However, it must be emphasized that these findings are preliminary. Subject to data availability in the future, the analysis could be further elaborated by, for instance, including additional control variables in the models to account for other possible determinants of innovation in wind energy technologies (R&D expenditures, renewables policies directed to wind energy technologies, electricity consumption, electricity price, average wind speed, etc.). Another possible extension would be to control for papers’ quality by using the number of citations of a paper in the scientific literature, and to analyse networks of knowledge flows between OECD and non-OECD countries. Finally, it could also be interesting to use long-term memory models in order to explore the relationship between patenting dynamics in non-OECD countries and co-authorship between OECD and non-OECD countries.

REFERENCES

Cameron, C.A. and P.K.Trivedi (1986), "Econometric models based on count data: Comparisons and applications of some estimators and tests", *Journal of Applied Econometrics*, 1(1): 29-53.

Cho, C.C., M.W. Hu, and M.C. Liu (2010), "Improvements in productivity based on co-authorship: a case study of published articles in China", *Scientometrics*, 85(2): 463-470.

Choi, S. (2012), "Core-periphery, new clusters, or rising stars? International scientific collaboration among advanced countries in the era of globalization", *Scientometrics*, 90(1): 25-41.

Franz, P. (2010), "Knowledge spillovers as a central element in theories about knowledge-based regional development: Advancement in theory and obstacles for empirical research", *IWH-Discussion Papers*, (5):1-22.

Furman, J.L., M.K. Kyle, I.M. Cockburn, and R. Henderson (2006), "Public & private spillovers, location and the productivity of pharmaceutical research", Technical report, National Bureau of Economic Research.

Gao, X. J. Guan, and R. Rousseau (2011), "Mapping collaborative knowledge production in China using patent co-inventorships", *Scientometrics*, 88(2): 343-362.

Guan, J. and Z. Chen (2012), "Patent collaboration and international knowledge flow", *Information Processing & Management*, 48(1): 170-181.

Haščič, I., et al. (2010), "Climate Policy and Technological Innovation and Transfer: An Overview of Trends and Recent Empirical Results", *OECD Environment Working Papers*, No. 30, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/5km33bnggcd0-en.

Jaffe, A.B. (1989), "Real effects of academic research", *American Economic Review*, pages 957-970.

Jaffe, A.B. and M. Trajtenberg (2005), *Patents, citations, and innovations: A window on the knowledge economy*. MIT press.

Jaffe, A.B., M. Trajtenberg, and R. Henderson (1993), "Geographic localization of knowledge spillovers as evidenced by patent citations", *The Quarterly Journal of Economics*, 108(3):577-598.

OECD (2010), *Measuring Innovation: A New Perspective*, OECD Publishing, Paris. DOI: http://dx.doi.org/10.1787/9789264059474-en

Popp, D (2002), "Induced innovation and energy prices", *American Economic Review*, 92(1):160-180.

Thompson, P. and M. Fox-Kean (2005), "Patent citations and the geography of knowledge spillovers: A reassessment", *American Economic Review*, pages 450-460.

Zeger, S.L. and K.-Y. Liang (1986a), "Longitudinal data analysis for discrete and continuous outcomes", *Biometrics*, 42(1):121-130.

Zeger, S.L. and K.-Y. Liang (1986b), "Longitudinal data analysis using generalized linear models", *Biometrika*, 73(1):13-22.

ANNEX. MODELLING STRATEGY

The Poisson model

Poisson regression models provide a standard framework for the analysis of count data (Cameron et al., 1986 [1]). A random variable Y is said to have a Poisson distribution with parameter λ if it takes integer values $y = 0,1,2,\ldots$ with probability:

$$\Pr ob\left[Y = y_i \middle| x_i\right] = \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i\,!} \quad for\,\lambda > 0\; and\; i = 1,\ldots,N \qquad (3)$$

where x_i is a vector of covariates and $i = 1,\ldots,N$ indexes the N observations of a random sample.

The log-linear conditional mean function of the Poisson model $E\left[y_i \middle| x\right]$ and its variance

$\operatorname{var}\left[y_i \middle| x_i\right]$ can be shown to be equal: $E\left[y_i \middle| x\right] = \operatorname{var}\left[y_i \middle| x_i\right] = \lambda$.

Since the mean is equal to the variance, any factor that affects one will also affect the other. Thus, the usual assumption of homoscedasticity would not be appropriate for Poisson data. The likelihood function for N independent Poisson observations is a product of probabilities given by Equation 3. Taking logs and ignoring a constant involving $\log(y_i\,!)$, we find that the log-likelihood function is :

$$\log L(\beta) = \sum y_i \log(\lambda_i) - \lambda_i \qquad (4)$$

where λ depends on the covariates x_i and a vector of p parameters β through the log-link of equation $\log(\lambda_i) = x_i^{'}\beta$.

However, in practice, count data are often over-dispersed relative to the Poisson distribution. Since observed data will almost always display pronounced over-dispersion, researchers typically seek alternatives to the Poisson model, such as the negative binomial model.

The Negative Binomial model (NB or NegBin)

The negative binomial model (Cameron et al., 1986 [1]) serves as a functional form that relaxes the equi-dispersion restriction of the Poisson model. A useful way to motivate the model is through the introduction of latent heterogeneity in the conditional mean of the Poisson model. The conditional mean function of the negative binomial model can be written as follows:

$$E\left[y_i \middle| x_i, \varepsilon_i\right] = \exp\left(\alpha + x_i^{'}\beta + \varepsilon_i\right) = h_i\lambda_i \qquad (5)$$

where $h_i = \exp(\varepsilon_i)$ is assumed to have a one parameter gamma distribution, $G(\theta,\theta)$ with mean 1 and variance $1/\theta = \kappa$;

$$f(h_i) = \frac{\theta^{\theta} \exp(-\theta h_i) h_i^{\theta-1}}{}, \; h_i \geq 0, \theta > 0 \qquad (6)$$

After integrating h_i out of the joint distribution, we obtain the marginal negative binomial distribution,

$$\mathrm{Pr}ob[Y = y_i | x_i] = \frac{\Gamma(\theta + y_i) r_i^{\theta} (1 - r_i)^{y_i}}{\Gamma(1 + y_i)\Gamma(\theta)} \qquad (7)$$

where $y_i = 0,1,\ldots,\theta > 0$ and $r_i = \dfrac{\theta}{\theta + \lambda_i}$.

The latent heterogeneity induces over-dispersion while preserving the conditional mean, that is to say:

$E[y_i | x_i] = \lambda_i$ and $\mathrm{var}[y_i | x_i] = \lambda_i \left[1 + \dfrac{1}{\theta}\lambda_i\right] = \lambda_i [1 + \kappa \lambda_i]$, where $\kappa = \mathrm{var}(h_i)$.

The conditional variance of the negative binomial distribution exceeds the conditional mean. If the estimate of dispersion after fitting is not near 1, then the data may be over-dispersed if the dispersion estimate is greater than 1 or under-dispersed if the dispersion estimate is less than 1. Over-dispersion results from neglected unobserved heterogeneity. The Poisson distribution is a special case of the negative binomial distribution where $\kappa = 0$.

The negative binomial model is estimated using maximum likelihood, where the likelihood function for the negative binomial model is such that:

$$L(\beta | y_i, x_i) = \prod_{i=1}^{N} \mathrm{Pr}ob[Y = y_i | x_i].$$

When count data are analysed with a negative binomial regression model, a dispersion parameter is estimated. If the dispersion is 0, then a Poisson model could be more appropriate. On the contrary, if the dispersion parameter is significantly different from 0, based on the 95% Confidence Limits for this parameter, then the choice of a negative binomial model is well justified.

The Negative Binomial model using the GEE method

The generalized estimating equations (GEEs), introduced by Liang and Zeger (1986 [13] [14]), is a method of analysing correlated data that otherwise could be modelled as a generalized linear model. GEEs have become an important strategy in the analysis of correlated data. These data sets can arise from longitudinal studies, in which subjects are measured at different points in time, or from clustering, in which measurements are taken on subjects who share a common characteristic.

Using the GEE method to estimate our negative binomial model enables us to obtain robust standard errors for the negative binomial regression coefficients. The robust standard errors attempt to adjust for heterogeneity in the model.

In longitudinal research with more than two assessments, repeated measures of discrete, normal and non-normal outcome variables can be analysed using GEE. GEE and HLM have important advantages that include the possibility (a) to model non-normal outcome variables, (b) to account for individual differences in behaviour change, and (c) to model the variance-covariance structure of the longitudinal data. We discuss each of these advantages in turn.

GEE is a particularly useful tool for longitudinal group comparisons with non-normal outcomes and multiple post-intervention assessments. In contrast to the common fixed effects models (e.g., ANOVA), GEE estimates population-averaged model, using an extension of the quasi-likelihood approach. Quasi-likelihood makes few assumptions about the distribution of the dependent variable and, for that reason, is applicable to a wide variety of non-normally distributed outcome variables. The only requirement involves the specification of the mean-covariance structure. GEE uses an iterative procedure for the development of an estimator whose error has a mean of zero and is asymptotically multivariate Gaussian. However, this requires that missing observations be missing at random. In GEE, the data are modelled by specifying the appropriate distribution family for the dependent variable (e.g., Poisson, negative binomial). If the data are not normally distributed, GEE is likely to yield considerably more test-power compared to repeated measures ANOVA with normalized variables.

However, it is important to ensure that the specified distribution family provides a good fit for the dependent variable. A Poisson model is quite restrictive in its assumptions and may not be appropriate for most count measures. Falsely specifying a Poisson distribution may produce misleading results and indicate significant effects that may not apply to the true distribution of the outcome. In general, the significance achieved with the specification of a particular distribution family or correlation structure is not a valid indicator of the appropriateness of the model. Even when working with normally distributed outcome variables, GEE might be preferred over classical ANOVA models because GEE treats individual change as a random variable. The advantages of a mixed design with a random individualized change variable and a fixed treatment group effect factor may be seen in a potentially increased test power. Further, GEE analyses are flexible in that they allow specifying the within-group correlation structure for the panels.

The Zero-Inflated Negative Binomial model

One frequent manifestation of over-dispersion is that the incidence of zero counts is greater than expected for the Poisson distribution and this is of interest because zero counts frequently have special status. Therefore, when there is a large number of zeros in the sample, we can estimate the zero-inflated variant of the negative binomial model in which the count process and the binary process are modelled separately. In this case, the proportion of zeros ϕ is added to the probability distribution while reducing other frequencies by a corresponding amount with the following mean and variance:

$$E[y|x_i] = \lambda(1-\theta) \text{ and } \mathrm{var}[y_i|x_i] = (1-\phi)(\lambda + \lambda\theta^2).$$

The proportion ϕ is parameterized by a logistic transformation of $z'\gamma$. The two parameters vector β and γ are to be estimated.

An inflation model, which includes logit coefficients, is used for predicting excess zeros. In our case, the explanatory variables are the total number of publications (TOTAL_PUBLI) and the total number of patents (PAT_TOTAL). In order to determine which model performs best, some tests can be conducted. For that purpose, the Akaike and the Bayesian information criterion (AIC, BIC) can be used.