Scientometrics (2013) 97:129–148
DOI 10.1007/s11192-013-1008-6

A concept for inferring 'frontier research' in grant proposals

Marianne Hörlesberger · Ivana Roche · Dominique Besagni · Thomas Scherngell · Claire François · Pascal Cuxac · Edgar Schiebel · Michel Zitt · Dirk Holste

Received: 2 December 2011 / Published online: 3 April 2013

Abstract This paper discusses a concept for inferring attributes of 'frontier research' in peer-reviewed research proposals under the popular scheme of the European Research Council (ERC). The concept serves two purposes: firstly to conceptualize, define and operationalize in scientometric terms attributes of frontier research; and secondly to build and compare outcomes of a statistical model with the review decision in order to obtain further insight and reflect upon the influence of frontier research in the peer-review process. To this end, indicators across scientific disciplines and in accord with the strategic definition of frontier research by the ERC are elaborated, exploiting textual proposal information and other scientometric data of grant applicants. Subsequently, a suitable model is formulated to measure ex-post the influence of attributes of frontier research on the decision probability of a proposal to be accepted. We present first empirical data as proof of concept for inferring frontier research in grant proposals. Ultimately the concept is aiming at advancing the methodology to deliver signals for monitoring the effectiveness of peer-review processes.

Keywords Frontier research · Peer-review · Bibliometric indicators · Lexical analysis · Statistical modelling

M. Hörlesberger (✉) · T. Scherngell · E. Schiebel · D. Holste
AIT Austrian Institute of Technology GmbH, Donau-City-Strasse 1, 1220 Vienna, Austria
e-mail: marianne.hoerlesberger@ait.ac.at

I. Roche · D. Besagni · C. François · P. Cuxac
CNRS, Institut de l'Information Scientifique et Technique, 2 Allée du Parc de Brabois, CS 10310, 54519 Vandoeuvre-les-Nancy, France

M. Zitt
INRA, Rue de la Geraudière, BP 71627, 44316 Nantes, France

M. Zitt
Observatoire des Sciences et des Techniques (OST), 93 Rue de Vaugirard, 75006 Paris, France

Introduction

Are peer review process on the one hand and assessment with bibliometric/scientometric methods for scientific output or grant applications on the other hand opponents or a successful team in scientific evaluation processes? This contribution reveals in which the combination of both methods, the peer review process and a bibliometric process, would support each other. Peer review process is a well-established for assessing scientific output. The output can be scientific literature in scientific journals, or conference contribution, or in research funding schemes as it is in the European Framework Programme, for grants in the programme of the European Research Council (ERC). The need for monitoring the effects and the implicit orientation peer review selection processes is subject to current research activities (Hojat et al. 2003; Sweitzer and Cullen 1994; Bornmann and Daniel 2008; Marsh et al. 2008).

As a result of the development in ICT scientific output is stored in databases. Scientists and (computer scientists, of physicists, or linguists, for instance) have increasingly addressed research in this field in the last decades. The research field informetrics, scientometrics, bibliometrics has risen activity, which emerges for instance in journals such as "Scientometrics", and in turning up of pertinent conferences show. Consequently scientometric evaluation has been witnessing a significant attention in the rising need to get a grip on science output and efficiency (see e.g., van den Besselaar and Leydesdorff 2009; van Noorden 2010). Science Citation Index, the Journal Impact Factor, *H*-Index etc. (Hirsch 2005; Jin 2006; Egghe 2006; Jin et al. 2007) are applied for the evaluation of universities, scientists, etc. although Eugene Garfield, one of the developers of them, and many of his colleagues have had to face various criticisms concerning measuring science. Nevertheless the huge amount of science production and the availability of electronic hardware and software offer this new research field.

Funding organisations such as the European Framework Programme, the European Research Council (ERC), and the national funding agencies are confronted with the evaluation of in some cases with a huge amount of proposals. Especially the ERC has to deal with a very high amount of applications for grants. Therefore the ERC launched a call for developing a model/concept for quantitative evaluation of grant applications. The basic idea was to combine the two methodologies, the peer review process and a quantitative assessment with scientometric/bibliometric indicators. The quantitative assessments with bibliometric methods could provide a first ranking with the output values of a well-defined set of indicators which can be a helpful input for peer reviewers. A closer inspection even showed a mutual dependence of peer-review and scientometric-based evaluation (Zitt and Bassecoulard 2008).

The idea of this contribution was to develop a bibliometric model for quantitative assessment of proposals of grant applicants. The bibliometric model needs indicators which are developed based on the definition of "frontier research" as the ERC understand it.

The paper discusses first "frontier research" as the ERC with its High-Level-Expert Group (HLEG) has defined it and its and transformation into bibliometric indicators. After that the data basis is presented. The core objective is the development of the concept which is feed by four/five indicators, which measure the different aspects of "frontier research". Therefore the development of these different indicators is introduced. A discrete choice model (DCM) is transformed into our bibliometric model for which the developed indicators are the input value. The last chapter discusses the obtained results.

The paper reveals that the proposed concept and model can well detect outliers assesses the proposals regarded the applied indicators. In this contribution we concentrate more on the evaluation of proposals especially in the framework of the ERC. The work for this paper is guided by the scheme of the ERC for the evaluation of grant applicants. Nevertheless the developed concept can be transferred to other analogical evaluation processes in science.

Background

Intense research is on-going for the identification of particular shortcomings and the further improvement of the significance of bibliometric evaluation and the development of science and technology indicators (Gorraiz and Schiebel 2008). Bibliometric and scientometric methods (carry strengths in that they are precisely defined and reliable, objective inter-subjective, efficient, and often need no human intervention. On the other hand, their weakness comes mainly in terms of limits of interpretation, applicability, confounding factors, and predictive validity (see e.g., Adam 2002; van Noorden 2010).

Moreover, many questions related to the attributes of frontier research are accessible by evaluating the patterns of the related publications activity. It is possible to estimate the impact of research results by evaluating the forward citations, the recentness of the work can be approximated by the timely distribution of the backward citations, the uniqueness of the work can be compared by determining the "market share" of the researcher in question in the related field of activity (Klavans and Boyack 2006, 2008; Boyack and Klavans 2010; Czerwon and Glänzel 1995).

Table 1 lists selected examples of scientometric indicators that are performance-centred and use easy-to-measure (in principle) volume data. The bottom row of Table 1 opens towards positioning and network indicators, some of them used for complementing performance-oriented indicators or identifying topics were indicators are calculated (such as co-authorship indexes, co-citation research fronts, etc.). Other types of indicators include data from curriculum vitae, e.g., age, gender; teaching-oriented measures; or referee scores. The multiple issues of socio-economic relevance are far from being addressed with mature indicators (e.g., extrapolation of trends on previously attracted, requested funds).

It is clear, that the identification of "frontier research" by means of bibliometric analysis seems a very ambitioned endeavour. However policy maker and research

Table 1 Selected classical scientometric indicators for measuring scientific performance

Indicator	Measurement	Interpretation options
Authorship	Number of publications or co-publications in specified sources, wide or selected coverage; world "market share" of publications[a]	Research output, productivity
Citation and impact at the journal level: impact factor and variants	Average citations per publication at the journal level[a]	Reputation of scientific journals
Citation and impact	Number of citations, world "market share" of citations, actors' impact factor, relative citation ratio, citation profile analysis[a], chains of influence	Research influence, international impact
Publication-citation	*h*-index, *g*-index, and related measures	Research productivity and impact
Online access	Number of times a paper is accessed online in some time period *T*	Indicators of use are, e.g., global spread, attention in scientific community and beyond
Network properties	Social network parameters (applicable, e.g., to co-working, co-publication, citation, word contents)	Beyond stand-alone, reflect system properties and influence in network interconnectedness and speed of information exchange

[a] With various conventions of counting/normalization

programme agencies have to face the developments in this field of informetrics/scientometrics/bibliometrics. The increasing challenges regarding the management of huge amount of proposals in our case and the possibility for some kind of impartial approach on the one hand combined with peer review on the other can be the chance for a new level of assessing science. Models of this kind may serve to verify decisions, deliver support data efficiently, or hint at biases of the review process (Juznic et al. 2010). Broadly, one can distinguish reviews in submitted manuscript (journal) reviews and reviews of proposed research projects. While the nature and objectives of the former centres on co-authorship, selection, improving quality of published research, etc., the latter focuses on the individual investigator, allocation of resources to inherently risky, speculative projects, etc. Differences between reviews of journals and research projects are particularly evident in different expectations on predictive validity of peer-review and, consequently, the choice of indicators tailored to the underlying strategy, mission and policy of publishers' responsibility funding bodies to establish interpretable and useful cause-effect relationships.

Frontier research transformation into bibliometric indicators

The definition of "frontier research" was developed by the HLEG, and described in the documented in (EC 2005). The four most important aspects for our work are summarized here:

- *Frontier research* stands at the forefront of creating new knowledge and developing new understanding. Those involved are responsible for fundamental discoveries and advances in theoretical and empirical understanding, and even achieving the occasional revolutionary breakthrough that completely changes our knowledge of the world. This aspect can be addressed by measurements of the indicator "NOVELTY"[1] (in two specifications, "TIMELINESS" and "SIMILARITY").
- *Frontier research* is an intrinsically risky endeavour. In the new and most exciting research areas, the approach or trajectory that may prove most fruitful for developing the field is often not clear. Researchers must be bold and take risks. Indeed, only researchers are generally in a position to identify the opportunities of greatest promise. The task of funding agencies is confined to supporting the best researchers with the most exciting ideas, rather than trying to identify priorities. This aspect can be addressed by measurements of the indicator "risk", here the personal risk of a scientist when he/she step out of his/her science environment.
- The traditional distinction between 'basic' and 'applied' research implies that research can be either one or the other but not both. With frontier research researchers may well be concerned with both new knowledge about the world and with generating potentially useful knowledge at the same time. Therefore, there is a much closer and more intimate connection between the resulting science and technology, with few of the barriers that arise when basic research and applied research are carried out separately. This aspect can be addressed by measurements of the indicator "PASTEURESQUENESS" (Following Stokes's concept of Pasteur's Quadrant).
- *Frontier research* pursues questions irrespective of established disciplinary boundaries. It may well involve multi-, inter- or trans-disciplinary research that brings together researchers from different disciplinary backgrounds, with different theoretical and

[1] The indicators are introduced and discussed in the following. Here they are firstly named.

conceptual approaches, techniques, methodologies and instrumentation, perhaps even different goals and motivations. This aspect can be addressed by measurements of the indicator "interdisciplinarity".

The following table (Table 2) gives an overview of the transformation of frontier research to bibliometric indicators. The development and specification of each indicator is discussed afterwards.

Data

The introduced model and the necessary indicators have been developed and tested on data of grant application to the ERC. The Table 3 gives an overview, by ERC call type and year, of the number of submitted proposals, awarded grants, and total allocated budget.

Table 2 Relation between the ERC definitions of frontier research, key attributes, bibliometric indicators, and approach to implement the extraction of attributes

Frontier research	Key attribute	Bibliometric indicator	Approach
Frontier research stands at the forefront of creating new knowledge	Novelty of the proposed research	TIMELINESS	Backward cited references
		SIMILARITY	Diachronic cluster analysis based on textual information
Frontier research is an intrinsically risky Researchers must be bold and take risks	Risk of the investigator through establishing scientific independence and/or taking on a new research field	RISK	Originality of the proposed research based on reference information of the proposal and principal investigator
Frontier research may be concerned with both new knowledge and with generating potentially useful knowledge Frontier research may take into account basic research and applied research.	Applicability (entrepreneurial principal investigator; proposed research)	PASTEURESQUENESS	Applicability of the expected results
Frontier research should involve multi-, inter- or trans-disciplinary research that brings together researchers from different disciplinary backgrounds, with different theoretical and conceptual approaches, techniques, methodologies and instrumentation, perhaps even different goals and motivations.	Science of interdisciplinary nature	INTERDISCIPLINARITY	Diversity reflected of the proposal on related panels (i.e., ERC-defined scientific disciplines) other than the own "home" panel based on textual information

Source definition: EC (2005); indicator: own data

Table 3 Number of proposals submitted and grants awarded by the ERC in 2007–2009

ERC grant (year)	Total budget (m €)	Number of proposals submitted	Total number of grants awarded	Number of grants awarded (in PE and LS)
SG (2007)	335	9,167	299	242
AG (2008)	553	2,167	282	198
SG (2009)	325	2,503	244	187
AG (2009)	515	1,584	244	202

Source ERC (2011); since 2009, SG and AG grants are awarded annually

SG Starting Grant, *AG* Advanced Grant

This paper looks at the scientometric evaluation of research project proposals in the following way:

- From a grant point of view, it focuses on proposals submitted to ERC in the scientific domains "Physics & Engineering" (PE) and "Life Sciences" (LS).[2] Scientists from all over the world, who are intending to work with a host institution based in a EU Member State or associated country, can compete for two different types of grants: Starting Grants (SGs) for investigators with 2–12 years of experience after their PhD at the stage of starting or consolidating their independent research team; and Advanced Grants (AGs) for already established investigators with at least 10 years of experience and significant research achievements (Antonoyianakis et al. 2009).
  Grants are to support pioneering, far-reaching research endeavours, combine high risk/high impact potential, break established disciplinary boundaries, or explore new productive lines of scientific enquiry, methodology or techniques.
- From a methodological point of view, it complements the standard approach to scientific excellence and specifically takes into account textual features related to the content and quality of '*frontier research*' (EC 2005) present in individual research proposals for awarding ERC grants to young or senior investigators (ERC 2008).

Due to the explicit formulation of the ERC's understanding of frontier research and its strategic importance for the funding scheme, ERC grants provide a suitable test-bed for content analysis/text-mining and modelling in the field of scientometric respectively bibliometric evaluation (Yoon et al. 2010). The primary interest is the extent to which research proposal comply with some attributes of frontier research and the influence of these attributes on the selection of awarded grants.

An external bibliographic database is also employed. It supplies the needed corpora of bibliographic records extracted by queries derived from the description of each considered ERC panel. The multidisciplinary bibliographic database PASCAL is used to provide a broad multidisciplinary coverage of more than 20 million records resulting from the

[2] PE (LS) holds ten (nine) main and ~170 (100) subcategories. The third domain "Social Sciences & Humanities" is not considered as it is expected to differ in terms of publishing, citation behaviour, and other features from those observed in PE and LS (e.g., national/regional orientation, less publications in form of articles, different theoretical 'development rate', number of authors, non-scholarly publications), which make it less assessable for approaches developed for the natural and life sciences (Nederhof 2006; Juznic et al. 2010).

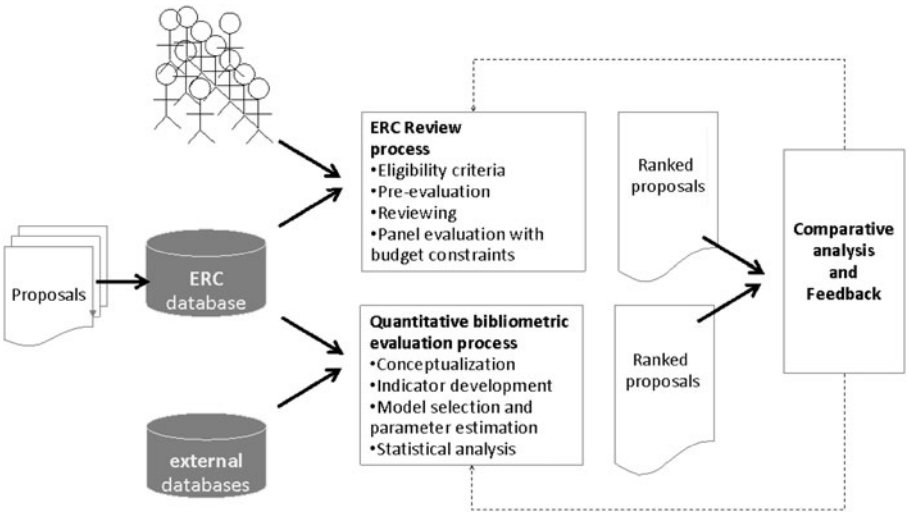

Fig. 1 The overall approach of for quantitative evaluation of the peer review process is shown in *two main branches*. The *upper branch*, which shows the classical review process for ranking and subsequent selection of positive evaluated proposals, is complemented by *second branch* with bibliometric model based ranking and evaluation of proposals. The quantitative process (*lower branch*) is compared with the qualitative peer review process for analyzing the influence of attributes of frontier research extracted from proposal data

analysis of the scientific and technological international literature that is published predominantly in journals and conference proceedings.

The research group (the co-authors) has concentrated their work here on the data coming from 2009 Call–Starting Grants. In the following the overall concept and the announced indicators are introduced and discussed.

Overall concept

The mode and indicators were developed based on the definition of frontier research (in general and regarding indicators for a specific field) as mention above. For evaluating the developed indicators and the output of the discrete choice model (DCM) the ranking of the proposals in question of the peer review process severed as benchmark (Fig. 1).

The relation between a sought quantitative model and the above definition of frontier research is made transparent through the correspondence between each identified key attribute and its indicator (Fig. 2).

Although each indicator has a clear description, quantification and interpretation, a faithful representation of frontier research (in a specific field/discipline) needs to combine them, which is implemented in form of a statistical model. We note that the notion of 'revolutionary breakthrough' (cf. Table 3) is practically inaccessible by scientometric and textual methods alone. Here two indicators capture different albeit related aspects of the research activity in question: the "timeliness" (one aspect of novelty) of the knowledge-base explicitly used by the author and the "similarity to emerging research topics" (another aspect of novelty) of the proposed research project inferred through the dynamic change of the scientific research landscape pertinent to this discipline.

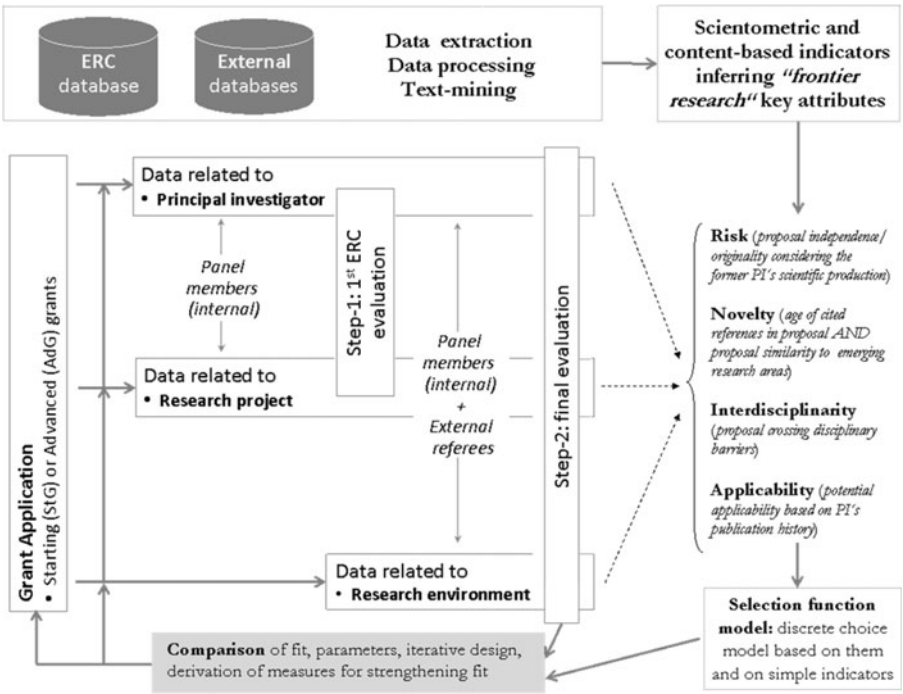

Fig. 2 The core concept for the quantitative (data-driven) evaluation is illustrated in three different categories. The first categorizes the incorporation of data related to the background of the principal investigator, which are captured in two indicators (RISK, NOVELTY). The next categorizes data related to the research project proposal, which reflected in tow other indicators (INTERDISCIPLINARITY, APPLICABILITY). The last category describes data related to the research environment of the work of the principal investigator and the proposal, which constitute the model environment of RISK and NOVELTY. The statistical model integrates and weights the influence of each parametrized indicator and allows for differential effects of attributes of frontier research

In computing indicators, an initial step identifies from a corpus of grant application relevant scientometric data (e.g., publications, citations, patents) and content data (e.g. text-strings, keywords) bearing relevance to frontier research, extracts and subjects them to data mining. In a subsequent step, actual indicators are computed and subjected to the model for comparison between peer-review panel and statistical model outcomes. Finally, model analysis and validation refine in a last step the performance of the model's usability. The following sections describe indicators, the model and proof of concept demonstration in more detail.

The indicators of frontier research

TIMELINESS

The first indicator, illustrated in Fig. 3, is based on citation analysis and used as one proxy to capture the "novelty" of a proposal through the bibliographic references cited within. The basic assumption is that the more recent references are, the more likely the work is at

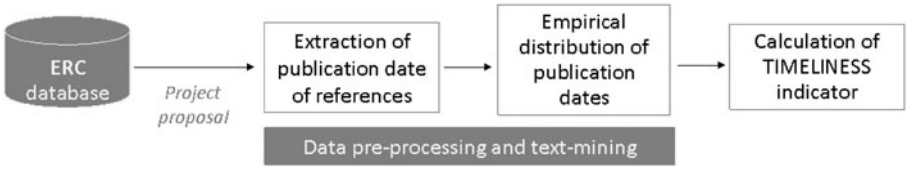

Fig. 3 The core bibliometric concept of the TIMELINESS indicator. The publication year of each reference cited by the applicant is extracted and employed in comparison with the year of the application submission

the cutting edge of the science. Therefore, the bibliometric concept focuses on references and the time elapsed since their publishing at the time of the proposal submission. TIMELINESS is obtained by considering for every reference in the proposal the relative difference in years between its publication date and the year of the application.

Stated references of the proposal are considered appropriate because not only do they relate directly to the project, but they constitute the knowledge base on which the proposal is built. Citation studies of obsolescence are conducted from two different perspectives: a diachronous obsolescence examining the citations received by a scientific publication within a particular time period; and a synchronous obsolescence examining references cited in a select set of documents at one point of time (Gupta 1997). Here we use the latter perspective where the set of documents is limited at each proposal.

After identifying references and extracting publication dates in actual texts, the indicator can be calculated from the set of values. The simplest way to obtain TIMELINESS is to use the arithmetic mean or median. Due to the fact that it may be influenced by statistical outliers, other statistics, re-sampling methods (e.g., bootstrapping) and comparison with known theoretical distribution are considered too to quantify the distribution.

For the aging of citations to scientific publications can depend on the field under study (Glänzel and Schoepflin 1995), one caveat is that indicators obtained for different panels may only roughly be comparable. As it describes the relative novelty of the proposal based on the set of references to the proposed research, TIMELINESS takes a narrower focus on 'novelty'. Its advantages (transparency, homogeneity, interpretation) are balanced by the limitations of presumably a short statistical basis and the fluctuations of behaviour expected among principal investigators. The novelty towards the state of the art is captured by the indicator SIMILARITY that is defined next.

SIMILARITY

The next indicator, illustrated in Fig. 4, operates a content analysis approach and is used to infer the "potential novelty" of a proposal. The core bibliometric concept rests on two pillars, the investigation of the on-going research in a field whose scientific perimeter is determined from the sub-panels describing a panel of Physics & Engineering and Life Science chosen in the ERC panel list on the one hand, and on the proposals applied to this field on the other hand. The investigation of the field is based on the construction of a "publication landscape" where increasing and emerging topics are identified (see also Chen 2005; Shibata et al. 2009; Small 1973; Srinivas and Viljamaa 2007).

To this end, raw data are extracted from external bibliographic databases (for international scientific and technological literature) by applying a query derived from the description of each considered ERC panel. The multidisciplinary bibliographic database PASCAL is used. Each PASCAL record is indexed, either manually by scientific experts or automatically based on a content analysis, by both keywords and thematic categories from

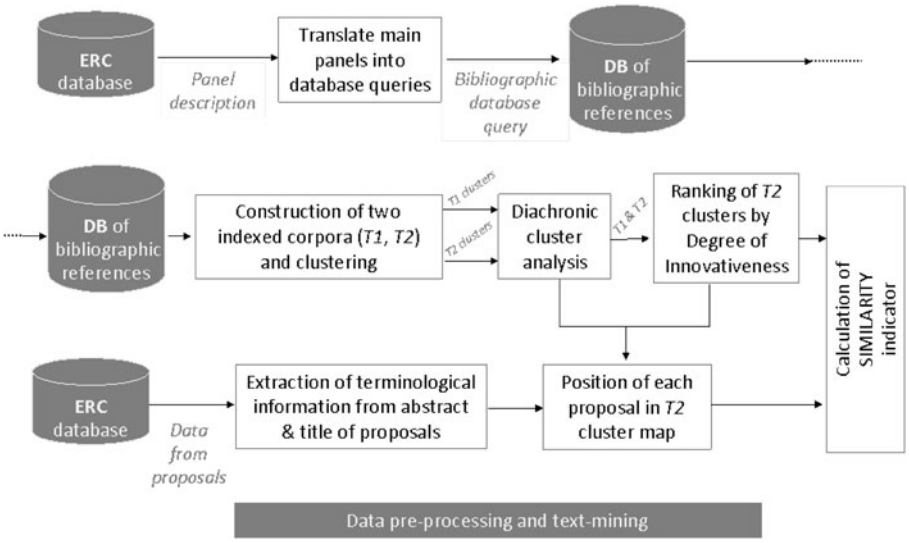

Fig. 4 The core approach for the SIMILARITY indicator. In the *upper branch* two sets of bibliographic records are produced, indexed, clustered and diachronically analysed in order to rank the obtained clusters by their innovativeness degree. In the *lower branch* the terminological information existing in each project proposal is extracted and employed to position it in the cluster map

a classification scheme. Nevertheless, the utilisation of not indexed information sources is also possible on the condition of a previous stage of text mining, to allow identifying the discriminating terminological information associated to each record. The SIMILARITY indicator is based on these indexing keywords.

Then, a clustering algorithm is used to obtain a cluster map which groups similar references on the basis of related keywords and represents the "publication landscape" corresponding to a considered ERC panel. The applied clustering tool applies a non-hierarchical clustering algorithm, the axial K-means method, coming from the neuronal formalism of Kohonen's self-organizing maps, followed by principal component analysis (PCA) to represent the obtained clusters on a 2-D map (Lelu and François 1992; Lelu 1993). This step is using the software system STANALYST (Polanco et al. 2001), specifically devoted to scientific and technological information analysis.

A diachronic analysis of the clustering results is used to study the evolution of the "publication landscape" across two subsequent time windows (referred here after as *T*1 and *T*2), by considering the content of each cluster and the variation of its relative location in the two networks of clusters. In particular the upper limit of the most recent period (*T*2) is considered as the year in which the analysed proposals were submitted. In this step structural alterations of the network of clusters between the two time periods are identified and analysed by a scientific expert (Roche et al. 2008): the splitting or disappearing, the persistence or emergence of clusters as well cluster status changes (e.g., cluster evolution from the periphery of the cluster network at *T*1 toward a central position in *T*2) are investigated. Techniques of association rule extraction are applied to determine the cluster evolution analysis by assessing the relationships between clusters of the two periods, applying the fuzzy association rules (Han and Kamber 2001; Hand et al. 2001; Mahgoub et al. 2008) through the so-called "confidence index".

The objectives are: (1) to establish which clusters potentially carry innovative topics and to class the set of clusters by rank of their innovativeness, which is determined from a calculated "inheritance index" and (2) to apply a methodology to evaluate the potential novelty of proposals by considering their similarity with respect to the clusters with the highest innovativeness values. We define the "inheritance index" as a measure of the relationships between the clusters from two periods, named $T1$ and $T2$ by using association rules. We use the fuzzy association rules because our items because keywords of the clusters resulting from the clustering step have non-binary weight values.

Considering only the direct relationships between the clusters of the second period with those of the first one could generate a loss of information while reducing its global relationship with the first period. It is for that reason that, in this work, two different indexes are calculated:

- The inter-period index (*InterP*) evaluates the direct relationship between the two periods and measures for each cluster of $T2$ the minimum confidence value among its relationships with each cluster of $T1$.
- The intra-period index (*IntraP*) takes into account the comparison exclusively between clusters from $T2$. It allows us to verify on the one hand whether these clusters are strongly linked together and on the other hand if they have potential indirect relationships with $T1$, which would not have been detected with *InterP*.

The global value of the "inheritance index" is defined as the harmonic mean of *IntraP* and *InterP* indexes. Thus a $T2$'s cluster with an "inheritance index" close to zero means that both indices are low. This means that this cluster is weakly linked, directly and indirectly, to the clusters from $T1$ and that its innovativeness is high, the keywords representing it dealing with topics potentially carrying positive dynamic changes (Roche et al. 2011).

In order to determine the potential novelty of a proposal with regard to $T2$ clusters, a text mining approach is firstly applied to extract from any considered proposal the terminological information to get a characterization as discriminative as possible while representing its content as faithfully as possible. Each proposal is represented by a binary vector showing the presence of its indexing keywords. The methodology calculates, for each proposal, its similarity to the $T2$ clusters and, by considering those with which it is the most similar, determines its SIMILARITY indicator value.

Metaphorically this means that each proposal is positioned in an evolving landscape and receives a SIMILARITY value depending on: (a) its calculated similarity to the $T2$ clusters and (b) the innovativeness value of those to which it is the most similar. The basic assumption is that the closer a proposal is to clusters having got high innovativeness values, the more novelty it is likely to carry.

PASTEURESQUENESS

This indicator, illustrated in Fig. 5, is based on both patent analysis as well as journal classification (applied, theoretical) and used to infer the applicability of expected results of each proposal, by considering evidence for immediate or intended application. Input data are obtained from bibliographic databases and proposals. The term PASTEURESQUENESS originates from the definition of Pasteur's Quadrant (Stokes 1997), which describes scientific research/methods that seek both fundamental understanding and at the same time social benefit.

The neologism PASTEURESQUENESS originates from the formalism introduced by Donald Stokes (1997) who defined a two dimensions chart, the "Pasteur's Quadrant". It is a label given to a class of scientific research developments that both seek fundamental understanding

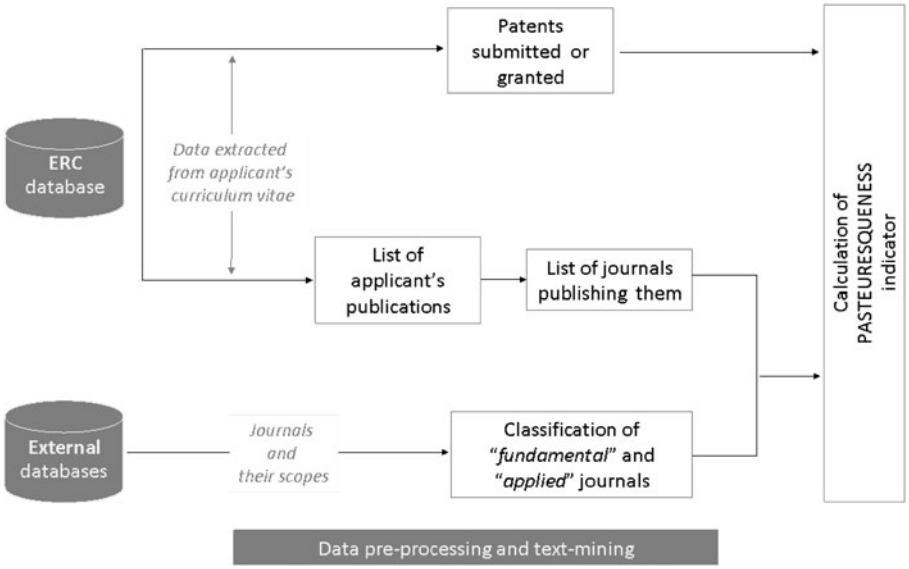

Fig. 5 The core approach for the PASTEURESQUENESS indicator. In the *upper branch* the data related to patents and to the journals in which the applicant's self-references were published are extracted. In the *lower branch* the categorization "applied versus fundamental" of these journals is obtained

of scientific problems, and, at the same time, seek to be eventually beneficial to society. The works of Louis Pasteur, a French chemist and physicist, pioneer of the microbiology, are thought to exemplify this type of method, which bridges the gap between "basic" and "applied" research. The Pasteur's Quadrant characterizes three distinct classes of research:

- *pure basic research*, illustrated by the work of Niels Bohr, early 20th century atomic Danish physicist;
- *pure applied research*, exemplified by the work of Thomas Edison, North-American inventor and businessman;
- *use-inspired basic research*, described as "Pasteur's Quadrant".

PASTEURESQUENESS consists of in all two indicators calculated from: on the one hand, patents granted or submitted with the participation of the principal investigator; information related to entrepreneurial involvement; and on the other hand the ratio of the principal investigator's articles published in journals of prescribed 'applicability'.

External data are used to determine the applicability degree of a journal scope. In the case of the PASCAL database, a categorization of journals in sub-fields was realized by experts of different scientific domains. This information is used to determine if a journal is applied or fundamental and, by analogy, this new categorization is employed to tag all the works published in the journal.

RISK

This indicator, illustrated in Fig. 6, is the second indicator that is based on citation analysis and used to infer the "individual risk" carried by the principal investigator in executing the proposed research. RISK is built by relying on references of proposals as well as published

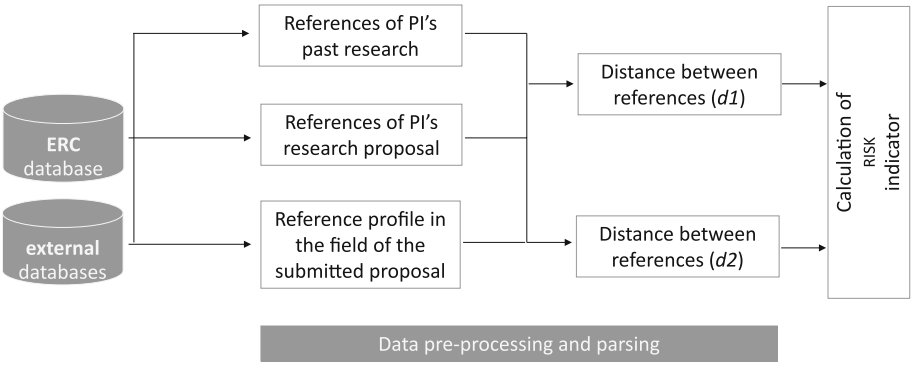

Fig. 6 The core approach for the RISK indicator. In the *upper branch* the data related to the cited references of a PI in his/her past research are analysed. The *middle branch* refers to the cited references in the grant proposal. The *lower branch* deals with the cited references in the whole filed of the submitted proposal

research papers as informative source for constructing a 'bibliometric research profile' (Hörlesberger et al. 2011). In comparing such a profile of a proposal to the research previously performed by the principal investigator, observed overlaps ($d1$) are used to categorize anticipated constant resp. aligned, modified or changed research directions. The incorporation of research contributed by peers inside the same research area ($d2$) is considered a second step and will be pursued in on-going work.

To this end, consider:

- all publications (n) of a principal investigator published in the past (defined as years prior to the year of submitting a proposal), extract cited references from publications and denote it by $\boldsymbol{R} = \{r_1, r_2, r_3, \ldots, r_m\}$, where r_i is the reference number i of the set $\boldsymbol{R}$ and occurs with a frequency (f_i); and
- all references of the proposal under consideration and denote it by $S = \{s_1, s_2, s_3, \ldots, s_p\}$, where s_i is the reference number i of the set $\boldsymbol{S}$ and occurs with a frequency (g_i).

If the principal investigator does not start out in a completely new research areas/direction, then there will be an expected overlap respectively intersection between the sets $\boldsymbol{R}$ and $\boldsymbol{S}$, e.g., $s_1 = r_2$, $s_2 = r_k$, $s_3 = r_i$, etc.

The correlation coefficient as a measure for linear statistical associations between R and S is used to calculate RISK. A positive value indicates bibliometric profiles that have more references in common whereas a negative value indicates the opposite. The basic assumption is that the lower the overlap between two reference profiles (past vs. proposed research), the more risk-affine is the proposal, because it is indicative of a change from previous pursued research.

INTERDISCIPLINARITY

The final indicator, illustrated in Fig. 7, is another indicator that is based on content analysis and used to infer self-consistently the presence and proportions of characteristic terminology that are associated with different ERC (home) panels, thereby revealing the interdisciplinary character of each proposal based on statistical properties of keywords across all panels.

It is built upon the basic assumption and previously successfully tested concept (Schiebel et al. 2010; Schiebel and Hörlesberger 2007) that the frequency of occurrence and distribution

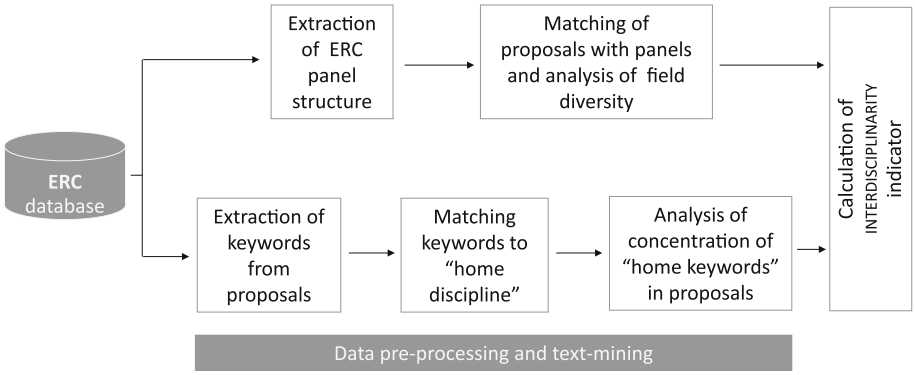

Fig. 7 The core approach for the INTERDISCIPLINARITY indicator. The *upper branch* uses pre-defined keywords selected by the principal investigator to match the proposal with scientific disciplines, while the *lower branch* uses keywords from proposals and compares the assigned panels of keywords (selected and fixed by the principal investigator when submitting the proposal) with their actual "home panels", i.e., panels that use keywords most frequently, which are self-consistently determined across all submitted proposals and panels

of discipline specific keywords in scientific documents can be used to classify and characterize disciplines. While the concept has been retained, the computation has been adopted to the grant scheme under study. The concept is consistent with the practice of bibliometric clustering, where the contents of each cluster (e.g., words and articles, or cited references and articles) are ranked by some index (e.g., TF–IDF) of specificity to the cluster.

The panel distributions of keywords are obtained in two different ways according to their descriptive nature:

- analysis of proposals based on an ERC pre-defined and finite set of keywords, selected by the principal investigator to describe the proposed research assigned to one major discipline out of about 30 in total in PE, LS, or SH; and
- analysis of proposals based on keywords freely selected by the principal investigator (e.g., free keywords, keywords from abstract information, or from proposal text, etc.) and comparison of the assigned home panel of keywords with the generally allocated panel of some proposal to reveal the intra- or inter-panel character of keywords for each proposal.

The underlying basic hypothesis is that the larger the proportion of inter-panel keywords, the more interdisciplinary is the proposal. To this end, each keyword is labeled according to its statistical frequency of occurrence across all PE or LS panels, filter are applied to distinguish relevant from irrelevant (i.e., panel unspecific) keywords, and the concentration of keywords with their assigned home panels is assessed to classify each proposals with respect to their inter-panel concentration.

The discrete choice model

Modelling the influence of indicators of frontier research for proposal acceptance

The set of bibliometric indicators and their captured aspects of frontier research are combined in a statistical model, with the aim to determine their influence upon the decision probability of a proposal to be accepted or rejected. Upon closer inspection, we further aim

to analyse the individual association of bibliometric indicators, i.e., which indicators show a comparatively higher (lower) influence on the decision probability.

From a strategic perspective, valid indicators of frontier research are expected to have a positive effect on the decision probability of a grant application, which has three possible outcomes: Type-A) above threshold and funded, Type-B) above threshold and not funded, and Type-C) below threshold. We use suitable methods from econometrics to address such questions with statistical rigor. In a first attempt, the decision model assumes a binary choice (types A/B vs. C) as they are the two central outcomes of the dependent variable: the rejection or acceptance of a project proposal. Denoting our set of observed project proposals by Y_i ($i = 1, \ldots, n$), we define our dependent variable by

$$Y_i = \begin{cases} 1 & \textit{proposal is accepted} \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

In econometric terms, we are dealing with a limited dependent variable model (see e.g., Greene (2003)), referring to situations where the dependent variable represents discrete alternatives rather than a continuous measure of activity. Specifically we borrow from the wide-spread class of discrete choice models, which are based on the in principle unobservable "utility", obtained from a specific choice among alternatives (Train 2009), which is to say the choice of a reviewer to accept or reject a project proposal. To this end, we define

$$X_i^{(k)} = \left(X_i^{(N)} \; X_i^{(R)} \; X_i^{(P)} \; X_i^{(I)} \; X_i^{(C)} \right) \tag{2}$$

where $\boldsymbol{X}_i$ is the joint vector of k ($k = 1,\ldots, K$) factors that may influence the decision probability of a proposal to be accepted, $\Pr(Y_i = 1)$. It comprises different vectors of variables that represent a specific type of frontier research: $X_i^{(N)}$is a vector of variables representing TIMELINESS as well as SIMILARITY, $X_i^{(R)}$ represents RISK, $X_i^{(P)}$ represents PASTEURESQUENESS, and $X_i^{(I)}$ represents INTERDISCIPLINARITY. Further we separate effects of these indicators from other intervening effects that are captured in model control variables ($X_i^{(C)}$). This yields the basic model as

$$\begin{aligned} \Pr(Y_i = 1) &= F(X_i^{(k)}, \beta) \\ \Pr(Y_i = 0) &= 1 - F(X_i^{(k)}, \beta) \end{aligned} \tag{3}$$

where $\boldsymbol{\beta}$ is the estimated k-by-1 parameter vector reflecting the impact of changes in $\boldsymbol{X}_i$ on the probability $\Pr(Y_i = 1)$, and $F(.)$ denotes the respective cumulative distribution function, which has to be chosen. It is common practice to use the logistic (logistic regression) where $F(.)$ is substituted with the logistic distribution function $\Lambda(.)$ so that the resulting logistic regression model reads as

$$\Pr(Y_i = 1) = \Lambda(X_i^{(k)}, \beta) = \frac{expX_i^{(k)}\,\beta}{1 - expX_i^{(k)}\,\beta} \tag{4}$$

where $\boldsymbol{X}$ is a set of k bibliometric and k control variables.

Proof of concept demonstration for a representative sample of ERC Starting Grants

In what follows we apply the model introduced above to a data set using 198 ERC starting grants of the year 2009, i.e., $i = 1, \ldots, n = 198$. These starting grants are composed of 41

proposals that have been selected by the ERC reviewers, while 157 of the proposals have been rejected. At this point, given our empirical situation, we are interested whether the dimensions of frontier research, TIMELINESS, SIMILARITY, RISK, PASTEURESQUENESS, and INTERDISCIPLINARITY, conceptualized and measured as introduced in the previous sections, indeed have influenced the reviewers' decision to select these 41 proposals.

Using our set of $i = 1,\ldots,n = 198$ proposals, we construct our binary dependent variable, and our independent variables including five measures for the five frontier research indicators, adding these variables to Eq. (4). At this point, we are interested to estimate the parameter vector $\beta = (\beta^{(1)}, \ldots, \beta^{(K)})$ that holds the information of how each variable influences the proposal decision probability. Thus the estimated parameters provide statistical evidence in the context of the proposed research question whether different attributes of frontier research—extracted from observed proposals and translated in scientometric terms—enhances the decision probability, and how these effects are statistically related to each other. We note that one well-known interpretation can be conducted in the form of probability odds, because from Eq. (4) it follows directly that

$$\frac{\Pr(Y_i = 1|X_{ik})}{1 - \Pr(Y_i = 1|X_{ik})} = \exp(X_i^{(k)}\,\beta). \tag{5}$$

Here $\exp(\beta)$ is the effect of the scientometric indicators on the odds, i.e., how a change of a specific variable affects the probability for a proposal to be accepted when all other variables are kept constant. The parameter estimation is based on standard Maximum-Likelihood techniques (see Greene (2003) for further details on the estimation procedure).

Table 4 presents the parameter estimates produced by Maximum-Likelihood estimation. The second column provides the respective parameter estimates, asymptotic standard errors are given in the third column. The modeling results point to interesting mechanisms playing a crucial role in the ERC evaluation process. The model produces significant estimates for the dimensions INTERDISCIPLINARITY and SIMILARITY, i.e., the ERC reviewers are indeed able to account for these frontier research dimensions in their decision finding process.

However, the parameter estimates for the remaining dimension are not statistically significant, that is TIMELINESS, RISK, and PASTEURESQUENESS, i.e., the model suggests that these dimensions—at least as measured in the current study—do not play a role in the ERC review process. From the perspective of the ERC, these results point to the necessity to improve the review process in a direction that also the insignificant dimensions are taken into account by the reviewers, and, by this, the model points to important conclusions for the ERC.

As demonstrated in the model specification, $\exp(\beta)$ is the marginal effect, and, thus, shows how a change in a specific exogenous factor affects the probability for a proposal to be accepted, when all other variables are constant. We can, thus, characterize the significant effects in more detail as follows: An increase of a proposal's INTERDISCIPLINARITY by 1 % increases the likelihood for proposal acceptance by a factor 1.14, holding all other variables constant. Further, an increase of a proposal's SIMILARITY to emerging research issues by 1 % increases the likelihood for proposal acceptance by a factor 1.69, holding all other variables constant.

Table 5 presents some model diagnostics. The Likelihood-Ratio test is statistically significant and confirms that adding the independent variables capturing our dimensions of frontier research increases the log-likelihood of the model, i.e. they significantly explain the variance of the dependent variable. The Hosmer–Lemeshow goodness of fit test that is

Table 4 Parameter estimates of the discrete choice model for the application of five bibliometric indicators

Variable	Parameter estimate	Standard error
Constant	−11.412*	0.433
Interdisciplinarity (β_1)	0.132*	0.023
Similarity (β_2)	0.524*	0.077
Pasteuresqueness (β_3)	0.077	0.121
Risk (β_4)	0.765	2.635
Timeliness (β_5)	−0.047	0.049

The independent variables are defined as given in the text (* significant at the 0.01 % level)

statistically insignificant confirms that the logistic link function is the right choice to explain the relationship between the dependent and the independent variables (see Greene (2003)). Further, various Pseudo R-squared measures (with a range [0,1]) underpin that the amount of variance that is explained by the independent variables is quite high.

Discussion

The above concept aims at advancing the development of quantitative methods for determining and examining the relationship between peer-review and decisions about research grant allocation in terms of attributes of frontier research:

- Can attributes of frontier research be faithfully represented and validly quantified to evaluate the grant allocation decision by bibliometric approaches?

The presented concept has focused on the ERC grant scheme but could be more broadly applicable depending on the mission, review process, attributes and correspondence of indicators for other grant schemes. Upon implementation it is intended to yield a scientometric model in which indicators can be used to show with statistical significance an effect on the decision probability for grant applications.

The concept introduced here utilizes information present in research proposals submitted to a grant agency and relates it to the bulk of information drawn from activities of the larger research community in a specific field. We have purposefully built upon both citation and textual analyses to address key attributes of frontier research. The full deployment of such characterization based on citation, keywords, or co-authorship, either in parallel or in complementary modules, leaves open many windows for research.

On the basis of the introduced indicators, immediate follow up questions address its usefulness:

- How does the model perform and what features does it reveal when applied to sample data extracted from specific grant application calls?
- How valid is the application of the model for the decision probability of research grant applications in terms of its model statistical properties?

The usefulness of the conceptualized and implemented indicators and corresponding model have been tested in a proof of concept approach, which is based on a representative sample of 198 proposals ($\sim$10 % of all submissions) for Starting Grants submitted in the year 2009. For data restrictions due to protection of intellectual property require the consent of applicants and hence a careful balance between dataset size to be gathered and the proposed benefit of the research of peer-review processes. The initial analysis of the sample SG2009 convincingly demonstrates such benefit in terms of a first proof of the

Table 5 Model diagnostic analysis for the application of the logistic regression with five indicators. All indicators show that the logistic regression model is valid for modelling the dependence of the selection probability

Likelihood ratio test	62.65*
Hosmer–Lemeshow Goodness of Fit	3.96 (Prob $> \chi^2 = 0.86$)
Efrons's *R*2	0.368
Cragg & Uhler *R*2	0.474
McFadden's Adj *R*2	0.310

* Significant at the 0.01 % level

indicator concept, model function selection and obtained results with statistical reliability as well as requirements of additional samples to study the generalizability of the direction of results and derived conclusions. In terms of the ERC review process, that is intended to explicitly select proposals reflecting frontier research, the results indicate that this aspiration holds when considering the INTERDISCIPLINARITY dimension of a proposal, and the SIMILARITY of a proposal to emerging research fields, that is even more important than INTERDISCIPLINARITY. However, the modelling results indicate that the ERC review process is not able to single out dimensions of frontier research that are related to TIMELINESS, RISK, and PASTEURESQUENESS.

Ultimately the concept shall advance the methodology to allow a grant agency to support the monitoring of the operation of the peer-review process from a scientometric perspective. In this context some ideas for future research come to mind that may increase the robustness of the results: *First,* controlling for additional variables that may influence decision probability, such as the number of citations or publications of the application, may be an essential addition to check the robustness of the results. *Second*, increasing the data set using proposals from other years and calls may be a valuable addition to control for time effects.

Acknowledgments The authors acknowledge the support that this work was partially funded by the Ideas specific programme of the EU's FP7 Framework Programme for Research and Technological Development (Project Reference No. 240765). The authors thank Helga Nowotny, Jens Hemmelskamp and Ulike Kainz-Fernandez of the ERC for stimulating discussions.

References

Adam, D. (2002). Citation analysis: The counting house. *Nature, 415*, 726–729.

Antonoyianakis, M., Hemmelskamp, J., & Kafatos, F. C. (2009). The European research council takes flight. *Cell, 136*, 805–809.

van den Besselaar, P., & Leydesdorff, L. (2009). Past performance, peer review and project selection: A case study in the social and behavioural sciences. *Research Evaluation, 18*, 273–288.

Bornmann, L., & Daniel, H. D. (2008). The effectiveness of the peer review process: Inter-referee agreement and predictive validity of manuscript refereeing at Angewandte Chemie. *Angewandte Chemie International Edition, 47*(38), 7173–7178.

Boyack, K. W., & Klavans, R. (2010). Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science (JASIST), 61*(12), 2389–2404.

Chen, C. (2005). Measuring the movement of a research paradigm. In *Proceedings of SPIE-IS&T, Visualization and data analysis*. V 5669, (pp. 63–76). San Jose.

Czerwon, H. J., & Glänzel, W. (1995). A new methodological approach to bibliographic coupling and its application to research-front and other core documents. *Proceedings of the 5th international conference on scientometrics and informetrics,* (pp. 7–10). River Forrest, Illinois, US.

EC (2005). Frontier research: The European challenge high level expert group report. European commission.
Egghe, L. (2006). Theory and practice of the g-index. *Scientometrics, 69*(1), 131–152.
ERC (2008). ERC Work Programme 2009.
Glänzel, W., & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature in the science. *Journal of Information Science, 21*(1), 37–53.
Gorraiz, J., & Schiebel, E. (Eds.) (2008). Excellence and emergence—a new challenge for the combination of quantitative and qualitative approaches. In *Book of abstracts of the 10th international conference on science and technology indicators Vienna, Austria*.
Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River: Prentice Hall.
Gupta, B. M. (1997). Analysis of distribution of the age of citations in theoretical population genetics. *Scientometrics, 40*(1), 139–162.
Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. San Francisco: Morgan Kaufmann Publishers.
Hand, D., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. Cambridge: The MIT Press.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *PNAS, 102*(46), 16569–16572.
Hojat, M., Gonnella, J. S., & Caelleigh, A. S. (2003). Impartial judgment by the "gatekeepers" of science: Fallibility and accountability in the peer review process. *Advances in Health Sciences Education, 8*(1), 75–96.
Hörlesberger, M., Holste, D., Schiebel, E., Roche, I., François, C., Besagni, D., et al. (2011). *Measuring the preferences of the scientific orientation of authors from their profiles of cited references*. Rome: ENID (European Network of Indicator Designers).
Jin, B. (2006). H-index: An evaluation indicator proposed by scientist. *Science Focus, 1*(1), 8–9.
Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin, 52*(6), 855–863.
Juznic, P., Peclin, S., Zaucer, M., Mandelj, T., Pusnik, M., & Demsar, F. (2010). Scientometric indicators: Peer-review, bibliometric methods and conflict of interest. *Scientometrics, 85*, 429–441.
Klavans, R., & Boyack, K. W. (2006). Quantitative evaluation of large maps of science. *Scientometrics, 68*, 475–499.
Klavans, R., & Boyack, K. W. (2008). U.S. Vulnerabilities in science and engineering. In *10th international conference on Science and Technology Indicators (S&TI), Vienne,* 17–20 September 2008, (pp. 86–88).
Lelu, A. (1993). *Modèles neuronaux pour l'analyse de données documentaires et textuelles*. PhD Dissertation: Université de Paris. 6.
Lelu, A., & François, C. (1992). Hypertext paradigm in the field of information retrieval: A neural approach. In *4th ACM conference on hypertext, Milano,* 30th November–4th December.
Mahgoub, H., Rösner, D., Ismail, N., & Torkey, F. (2008). A text mining technique using association rules. *International Journal of Computational Intelligence, 4*(1), 2008.
Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications—reliability, validity, bias, and generalizability. *American Psychologist, 63*(3), 160–168.
Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics, 66*, 81–100.
van Noorden, R. (2010). A profusion of measures. *Nature, 465*, 864–865.
Polanco, X., François, C., Royaute, J., Besagni, D., Roche, I. (2001). STANALYST: An integrated environment for clustering and mapping analysis on science and technology, In *8th international conference on scientometrics and informetrics, Proceedings Vol. 2, Sydney, Australia,* July 16–20, 2001, (pp. 871–873).
Roche, I., Besagni, D., François, C., Hörlesberger, M., Schiebel, E. (2008). Identification and characterisation of technological topics in the field of molecular biology, *Proceedings of the 10th international conference on Science and Technology Indicators (S&T I), Vienna*, 17-20 September 2008, (pp. 320–322).
Roche, I., Ghribi, M., Vedovotto, N., François, C., Besagni, D., Cuxac, P., Holste, D., Hörlesberger, M., Schiebel, E. (2011). Detecting domain dynamics: Association rule extraction and diachronic clustering techniques in support of expertise, *1st global TechMining conference, Atlanta, USA,* 14th September 2011.
Schiebel, E., & Hörlesberger, M. (2007). About the identification of technology specific keywords in emerging technologies: The case of "Magnetoelectronic", In *11th ISSI conference, Madrid, Spain,* 25–27 June 2007.

Schiebel, E., Hörlesberger, M., Roche, I., François, C., & Besagni, D. (2010). An advanced diffusion model to identify emergent research issues: the case of optoelectronic devices. *Scientometrics,*. doi:10.1007/s11192-009-0137-4.

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative study on methods of detecting research fronts using different types of citation. *Journal of the American Society for Information Science and Technology, V60*, 571–580.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science (JASIST), 24*(4), 265–269.

Srinivas, S., & Viljamaa, K. (2007). Emergence of Economic Institutions: Analysing the third role of universities in Turku, Finland. *Regional Studies, 42*(3), 323–341.

STOKES D. (1997). Pasteur's Quadrant, The Brookings Institution.

Sweitzer, B. J., & Cullen, D. J. (1994). How well does a journal's peer review process function? *Journal of the American Medical Association, 272*, 152–153.

Train, K. E. (2009). *Discrete choice methods with simulation* (2nd ed.). Cambridge: Cambridge University Press.

Yoon, B., Lee, S., & Lee, G. (2010). Development and application of a keyword-based knowledge map for effective research. *Scientometrics, 85*, 803–820.

Zitt, M., & Bassecoulard, E. (2008). Challenges for scientometric indicators: data demining, knowledge-flow measurements and diversity issues. *Ethics in Science and Environmental Politics, 8*, 49–60.