Journal of Informetrics 11 (2017) 282–298

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Toward an excellence-based research funding system: Evidence from Poland

Emanuel Kulczycki [a], Marcin Korzeń [b], Przemysław Korytkowski [b,*]

[a] Adam Mickiewicz University in Poznań, Poland
[b] West Pomeranian University of Technology in Szczecin, Poland

ARTICLE INFO

Article history:
Received 16 August 2016
Received in revised form 6 January 2017
Accepted 7 January 2017
Available online 26 January 2017

Keywords:
Metrics
Bibliometrics
Research evaluation
Excellence in research
Poland

ABSTRACT

This article discusses the metrics used in the national research evaluation in Poland of the period 2009–2012. The Polish system uses mostly parametric assessments to make the evaluation more objective and independent from its peers. We have analysed data on one million research outcomes and assessment results of 962 scientific units in the period 2009–2012. Our study aims to determine how much data the research funding system needs to proceed with evaluation. We have used correlation analysis, multivariate logistic regressions models and decision trees to show which metrics of the evaluation played a major role in the final results. Our analysis revealed that many metrics taken into account in the evaluation are closely correlated. We have found that in the Polish system, not all the collected data are necessary to achieve the main goal of the system, namely the categorization of scientific units in terms of their research performance. Our findings highlight the fact that there is a high correlation between performance in terms of publications and the scientific potential of a given scientific unit. We conclude with recommendations and a suggestion of a transition from a system in which the scientific units report all their metrics to a system in which they show only the most important metrics that meet the requirements of excellence in research.

1. Introduction

Performance-research funding systems (PRFSs) have been used as a science policy tool for the last three decades. The Research Assessment Exercise in the UK was launched in 1986. In the early 1990s, the assessment of research institutions began in Poland. Since then, many countries have introduced PRFSs and have embedded them in their national research systems (Bloch & Schneider, 2016; Hicks, 2012). The aim of developing and implementing PRFSs is to allocate funds to excellent institutions. To determine such institutions, research outcomes are evaluated *ex post*. The measurement methods which are used within such an evaluation can be divided into three categories (Aagaard, Bloch, & Schneider, 2015): peer review-based models, publication count-based models and citation-based models. The peer-review model is used in the UK's Research Excellence Framework. The other two methods, based on publication count and publication citations, are used in most other PRFSs (e.g. in the Czech Republic, Finland, Belgium/Flanders, Italy, Norway and Poland).

The performance of the scientific unit is constituted by various research outcomes, such as publications, projects, organised conferences and others. One could identify the most significant part of such a performance, but when we evaluate

* Corresponding author.
E-mail address: pkorytkowski@zut.edu.pl (P. Korytkowski).

http://dx.doi.org/10.1016/j.joi.2017.01.001
1751-1577/

a scientific unit's performance, we measure all of its activities. In the PRFSs, however, such an evaluation policy did not strongly emphasize whether the outcomes were actually desirable from the funders' point of view. Contrary to the concept of performance, the concept of "excellence in research" indicates that only some parts of the performance are desirable from the funders' point of view. Most researchers and stakeholders agree that excellence should be rewarded; yet, the challenge of how to define and quantify such "excellence" remains. (Arthur, 2015; Sunkel, 2015). In the UK, the PRFSs shifted from outcomes assessment in the Research Assessment Exercise to measurement of impact and outcome excellence in the Research Excellence Framework (Chowdhury, Koya, & Philipson, 2016). This transition was conducted in the PRFS in which the evaluation had been undertaken using the peer-review model. Such a model takes for granted that in the evaluation process only the most important outcomes should be assessed.

Hicks (2012) highlights that the complexity of PRFSs has increased over time, as they are dynamic systems being expanded by the addition of new indicators, such as h-index or altmetrics. When a system becomes more complicated, its indicators and metrics can be gamed (Rijcke, de Wouters, Rushforth, Franssen, & Hammarfelt, 2016), and it becomes even harder to legitimize such a model of research evaluation. Thus, as the authors of the "Leiden Manifesto" (Hicks, Wouters, Waltman, de Rijcke, & Rafols, 2015) suggest, the indicators should be regularly scrutinized and updated. In this way, changing and improving PRFSs always face challenges such as keeping a balance between peer review and metrics, universal and specific field models, considering all activities and only important ones and redistributing funding to the best performers and motiving all performers. This is possible to achieve in a variety of ways, such as differentiated publication counts (Schneider, Aagaard, & Bloch, 2014) or expert-based and citation-based ranking of publication channels (Saarela, Kärkkäinen, Lahtonen, & Rossi, 2016).

Several countries, such as Australia, Brazil, France, Italy and Poland, have developed national journal rankings in which all the disciplines are represented (Ferrara & Bonaccorsi, 2016; Haddow & Genoni, 2010; Vanclay, 2011). These rankings serve to indicate which publications 'do count' in a national evaluation. Publications form the greatest part of evaluated research outcomes. For instance, the Italian Research Evaluation assessment for the period 2004–2010 analysed almost 184,000 publications and 1000 other research outcomes (Ancaiani et al., 2015). The Czech PRFS evaluated over 97,000 publications and 10,000 other research outcomes for the period 2005–2009 (Good, Vermeulen, Tiefenthaler, & Arnold, 2015). The Polish PRFS evaluated over 184,000 publications and 182,000 other research outcomes for the period 2009–2012.

Even though publications are the most significant component in the evaluated outcomes, there are many other scientific activities that characterize well-performing institutions and determine the level of their productivity. Among these scientific activities are granted patents, acquired projects, artistic productions, scientific degrees awarded, scholarships received, editorial engagements, research infrastructures, scientific conferences organized and others. As many studies have also shown, the chosen category of scientific activity used within the PRFS has a significant influence on evaluating research outcomes (Abramo & D'Angelo, 2015; Ancaiani et al., 2015; Good et al., 2015; Sivertsen, 2015).

The current Polish PRFS is based on lessons from previous evaluations as well as consultations with the Polish scientific community. The Polish Ministry of Science and Higher Education declares that the main objective of this system is funding distribution to scientific units according to their scientific achievements. The quality of a given scientific unit is expressed as one of four categories: A+, A, B or C. The best scientific units, the A+ category, receive much greater financing, 150% of that allocated to the A category. Scientific units with the B category receive 70% of the funds allocated to the A category units and those with C category only 40% and only for half a year. The intention of this system was to strengthen scientific performance. This aim has been achieved, among others, by reducing the number of reported publications for the four-year reporting period.

During the consultations, many of the scientific units claimed that the evaluation system should take into account all aspects of their activities. In this way, the number of PRFS parameters has been inflated in Poland. This opened space for parameter optimization and the gaming of the system. On one hand, scientists criticize the Polish system for its excessive complexity, and, on the other hand, they often demand the addition of more parameters specific to their field of study or that their importance be bolstered by increasing the number of points awarded to a parameter.

In 2015, the Polish government published a strategy entitled the *Higher Education and Science Development Programme for 2015–2030* (MNiSW, 2015), which has established goals for current science policy. In this strategy, it is explicitly stated that the aim of the research evaluation system is to identify various aspects of excellence in research. In the previous strategy published in 2008, the concept of excellence was not used: performance was the very idea on which the research evaluation was focused. Moreover, in the current strategy, one can find that "in the parametric evaluation of scientific units, the quality of research outputs (especially prestigious publications and financially rewarded implementation) should be more important than quantity (. . .) In the evaluation, there should be acknowledged only the best research outcomes, and a number of the outcomes should not be just a simple multiplicity of the full-time employment equivalents" (MNiSW 2015, p. 22). Focusing only on the best and most important outcomes is relevant from the Polish science policy point of view. Poland's publication output is below the average European Union levels; for example, the percentage of highly cited publications for Poland is 6.36%, whereas the average EU level is 12.25% (Klincewicz & Szkuta, 2016). Thus, the improvement of Polish PRFS has to face the following challenge: if we agree that 'what gets measured gets done', then we should focus only on such types of research outcomes that achieve the goals of science policy.

Our study aims to answer the following question on the assessment of scientific institutions in Poland by using the collected datasets in the evaluation process: which metrics (and data) does a performance-based research funding system actually need? We have assumed that reducing the number of metrics or parameters would allow focusing on those types

of research outcomes that are important from the Polish science policy perspective. In this way, various incentives could be constructed that might influence researchers' behaviour. As in every complex system, this may have desirable but also undesirable consequences. However, at the beginning, we have to analyse whether such a parameter reduction is technically possible. Then we may discuss the social and political consequences of modifying the Polish PRFS.

The academic community in Poland awaits the next evaluation in 2017. Meanwhile, researchers, stakeholders and policy makers have been discussing new regulations that should shape the evaluation in 2021. One of the practical aims of the present paper is to provide arguments and assumptions for the fourth version of the Polish PRFS, which could better emphasize the idea of excellence in research.

The Polish PRFS has become a holistic dataset describing the four-yearly activities of all scientific units in Poland. The collected dataset gives us a unique opportunity to conduct a comprehensive statistical analysis. We have examined the system, the collected data and the evaluation results to improve the Polish PRFS and to reduce the drawbacks resulting from over-regulation of the system and from gathering too much irrelevant data.

This study is structured as follows: firstly, we present the framework of the most recent 'the third' iteration of the Polish PRFS and explain the metrics and parameters of the evaluation. Then, the data and methods are described, with the following section presenting the results, focusing on the existing relationships among the parameters within the groups of sciences. In the final section, we discuss the main findings.

2. The polish research evaluation system

In Poland, a preliminary assessment of research institutions was conducted in 1990. A year later, the State Committee for Scientific Research, which combined the role of a typical ministry of science and higher education with that of a funding agency, introduced a framework for the peer-review evaluation of all Polish research institutions. It was the first version of a Polish PRFS in which the institutions were categorized in terms of their scientific performance. The first categorizations were conducted from 1991 to 1998. Since then the, Polish PRFS has evolved in several cycles. The argument for changing the first version was a devaluation of assigned scientific categories, which resulted from the assignment of too many scientific units to the highest category. Thus, the second version of the Polish system was based on a parametric assessment to make the evaluation more objective and independent of its peers. In 1999, new regulations were introduced based on a parametric evaluation. This second version of the Polish PRFS shifted from a peer-review evaluation to a parametric evaluation in which the role of expert opinions was reduced. Moreover, at this time, the State Committee for Scientific Research started preparing a national scientific journal ranking to support the research evaluation system. Using the second version of the Polish PRFS, four cycles of evaluation were conducted: in 1999, 2003, 2006 and 2010. Each subsequent cycle was changed and improved from the previous cycle. The range of data, number of categories and definitions of parameters became more precise. In 2005, the State Committee for Scientific Research was merged into the Ministry of Science and Higher Education in Poland and continued the work on the Polish PRFS. The implementation of the third version was the result of establishing in 2010 a Committee for the Evaluation of Scientific Units – that is, an advisory board to the Minister of Science and Higher Education, which has since become the board responsible for national evaluation in Poland.

2.1. The 2013 evaluation

The current version of the Polish system has been cited in only a few studies (Aagaard, 2015; Hicks, 2012; Jonkers & Zacharewicz, 2016). Three scientific articles have been published in Polish and were devoted to the previous cycle of scientific unit evaluation in 2013, focusing on the analysis of results from the perspective of the humanities and social sciences (Antonowicz & Brzeziński, 2013; Kulczycki, Drabek, & Rozkosz, 2015; Sadowski & Mach, 2014). Kulczycki (2017) published a paper in English that describes the main components of the evaluation system in Poland and showed how the publication assessment system had been implemented. Koczkodaj, Kułakowski, and Ligęza (2014) analysed the consistency-driven pairwise comparison method that was used for building the final classification of scientific units in the 2013 evaluation.

The regulations for the 2013 evaluation were presented by the Ministry of Science and Higher Education in the Regulation of 13 July 2012 on the metrics and procedure for assigning scientific categories to scientific units. A scientific unit can be a unit within a higher education institution (most often a faculty), a basic research institute, an applied research institute or so-called "others", referring to companies performing R&D activities. The evaluation of these concerned four groups of sciences: social sciences and humanities (SSH), sciences and engineering (SE), life sciences (LS) and art sciences and artistic production (ASP).

For the 2013 evaluation, almost one million evaluation items were submitted by 962 scientific units representing 83,211 researchers for the period 2009–2012 (Skoczeń et al., 2014). An *evaluation item* is a single piece of data describing the scientific unit's research outcomes, such as a monograph, article, patent, project, artistic production, scholarship, research infrastructure or organized scientific conference. The evaluation results for the period 2009–2012 were published in July 2014.

2.2. Four metrics of evaluation

The evaluation items were assigned to one of four metrics: M1–scientific and creative achievements, M2–scientific potential (scientific strength), M3–material effects of the scientific activity and M4–other effects of the scientific activity. The metrics and component parameters were designed by an advisory group for the Ministry of Science and Higher Education in Poland, namely the Committee for the Evaluation of Scientific Units. As Table 1 below shows, each of the metrics (M1–M4) was applied in all the groups of sciences and types of scientific units. In each metric, there were various parameters that applied to all or just some groups of sciences. For instance, only scientific units from the SE and the LS groups could submit publications for evaluation in Metric 1, the parameter "Article in conference proceedings that are indexed in the Web of Science" (P1.1.4). Please note that the number of points allocated for a given parameter is not comparable with the number of points given to a parameter belonging to a different metric due to the separate pairwise comparison of aggregated results for each metric.

In each parameter, a scientific unit could obtain a specific number or range of points for an *evaluation item*. It is worth noting that in the Polish system, publications are counted only once in a single scientific unit, yet the entire count is applied when the authors work in different scientific units, meaning that a single publication written by four researchers from four different Polish scientific units can generate four *evaluation items*. The number of points obtained by a scientific unit across all the metrics serves to build the classification of the institution within the evaluation.

In Metric 1, the most important parameters are related to publications. The number of publications that a scientific unit can submit for evaluation is limited by two rules (Kulczycki, 2017). The first rule is expressed in the formula $3N - 2N_0$, where N is the arithmetic mean of the number of academic staff members who work in a given scientific unit during the period 2009–2012, while N_0 is the number of academic staff members who were not authors of any publication during the period in question. The other rule has limited the number of monographs that a scientific unit can submit for evaluation: for the SE and LS groups the limit is 10%, and for the SSH and ASP groups the limit is 40% of all submitted publications. A scientific unit can obtain points for publications in scientific journals according to the Polish Journal Ranking, the so-called "ministerial list of journals", which is prepared annually. The ranking consists of sub-lists, as follows: (1) *the A list* – journals listed in the Journal Citation Reports: the discipline-normalized five-year impact factor is translated into points; (2) *the B list* – journals without an Impact Factor and not indexed in the European Reference Index for Humanities (ERIH): the number of points depends on bibliometric and formal metrics; and (3) *the C list* – journals indexed in the ERIH: the number of points depends on the ERIH category (NAT, INT2, INT1). This metric includes reported outcomes with intellectual property rights such as patents and artistic works. The final value of the metric is normalized by dividing the sum of the points by N (the arithmetic mean of the number of academic staff members).

Metric 2 measures the potential (strength) of a scientific unit. This metric counts the number of authorisations for awarding academic degrees (one authorisation for one scientific discipline); the number of degrees awarded (broken down into employees and non-employees); the professional activities of scientific staff, such as memberships in international scientific organizations and on journal editorial boards and expert panels as well as publishing in journals indexed in the JCR or ERIH; and the status of nationally or internationally certified laboratories. In this metric, scientific units from the LS group could count in R&D projects. In the last evaluation, 82,519 items were assessed in this metric. In contrast to Metric 1, this metric isn't normalized by dividing the sum of points by N, favouring large scientific units that are able to receive more points because of a greater number of scientific staff members.

Metric 3 measures the financial flows of scientific units. Money received directly from the ministry on the basis of statutory funding was excluded from the evaluation. Only funds obtained from project competitions and from cooperation with industry or local authorities were taken into account. This metric was normalized in the same manner as Metric 1–that is, the sum of the points was divided by N (the arithmetic mean of the number of academic staff members). During the evaluation procedure, 91,431 items were analysed in this metric.

Metric 4 comprises all other activities that aren't taken into account by the other metrics, in other words, scientific activities that are hard to measure by metrics. Each scientific unit could submit up to 10 noteworthy activities, such as: conference organization, the application of research results, the dissemination of knowledge, activities that are of particular importance to national heritage or the development of culture and science. Items within Metric 4 were independently assessed by two experts who could assign from 0 to 100 points for the activities submitted by a given scientific unit. The final grade in this metric was the average of the experts' evaluation. In the last evaluation, 962 scientific units submitted 8797 activities for assessment in Metric 4.

The metric value was calculated as the sum of the parameter values. For instance, M2 for the social sciences and the humanities group (SSH) is the sum of parameters P2.1 + P2.2 + P2.3 + P2.4. Parameters P2.5–P2.7 are not taken into account in this group. Further, P2.1 is calculated again as the sum of P2.1.1 and P2.1.2 and so on for P2.2–P2.4. As was mentioned earlier, the final values of M1 and M3 are normalized by dividing the sum of points by N.

2.3. Scientific unit categorization

At the beginning of the research evaluation, all scientific units were assigned to Joint Evaluation Groups (JEGs) within the groups of sciences and particular type of scientific units; for example, faculties of philosophy were assigned to a single JEG designed for the units from higher education institutions from the SSH group, and institutes of the Polish Academy of

Table 1
The metrics and parameters of evaluation within the 2013 Comprehensive Evaluation of Scientific Units.

Metric 1: Scientific and creative achievements (M1)

ID	Parameter Name	Points	SSH	SE	LS	ASP
P1.1	**Journal articles**		+	+	+	+
P1.1.1	Article in a journal indexed in the Journal Citation Reports (JCR)	10–50	+	+	+	+
P1.1.2	Article in a journal indexed in the Polish Journal Ranking (the ranking indexes journals without an Impact Factor and not indexed in the European Index for the Humanities)	1–10	+	+	+	+
P1.1.3	Articles in a journal indexed in the European Reference Index for the Humanities (ERIH)	10–14	+	+	+	+
P1.1.4	Article in conference proceedings that are indexed in the Web of Science	10	–	+	+	–
P1.1.5	Article in a congress language** (other non-national journals)	4	+	–	–	+
P1.2	**Monographs**		+	+	+	+
P1.2.1	Monograph in a congress language**	25	+	+	+	+
P1.2.2	Monograph in Polish	20	+	+	+	+
P1.2.3	Chapter in a congress language**	5	+	+	+	+
P1.2.4	Chapter in Polish	4	+	+	+	+
P1.2.5	Edited volume in Polish (points for editing)	4	+	+	+	+
P1.2.6	Edited volume in a congress language** (points for editing)	5	+	+	+	+
P1.3	**Intellectual property rights**		-	+	+	+
P1.3.1	Proprietary rights of patent owned by the scientific unit	25	–	+	+	+
P1.3.2	Proprietary rights of patent owned by third party, employee is an inventor	15	–	+	+	+
P1.3.3	Trademarks, designs, utility models, semiconductor topography rights	10	–	+	+	+
P1.3.4	Plant variety rights	15	–	+	+	+
P1.3.5	Patent application	2	–	+	+	+
P1.4	**Artistic work**		-	+*	-	+
P1.4.1	Authorship of a major artistic work	25	–	+*	–	+
P1.4.2	Authorship of a minor artistic work	12	–	+*	–	+
P1.4.3	World premiere of a major artistic work	12	–	+*	–	+
P1.4.4	World premiere of a minor artistic work	6	–	+*	–	+
P1.4.5	Performance of a minor artistic work	20	–	+*	–	+
P1.4.6	Participation in a performance of an artistic work	10	–	+*	–	+
P1.4.7	Participation in the collective exhibition or in the restoration of works abroad	4	–	+*	–	+
P1.4.8	Participation in the collective exhibition or in the restoration of works domestically	2	–	+*	–	+
P1.O	**Contribution to the national defence**	****	-	+	-	-

Metric 2: Scientific potential (M2)

ID	Parameter Name	Points	SSH	SE	LS	ASP
P2.1	**Authorisations for awarding academic degrees**		+	+	+	+
P2.1.1	Authorisations for awarding DSc degrees	70	+	+	+	+
P2.1.2	Authorisations for awarding PhD degrees	30	+	+	+	+
P2.2	**Academic promotion of employees**		+	+	+	+
P2.2.1	Awarded PhD degrees count	2	+	+	+	+
P2.2.2	Awarded DSc degrees count	7,10	+	+	+	+
P2.2.3	Awarded professor titles count	10,14	+	+	+	+
P2.3	**Academic promotion of non-employees**		+	+	+	+
P2.3.1	Awarded PhD degrees to non-employees count	1	+	+	+	+
P2.3.2	Awarded DSc degrees to non-employees count	3	+	+	+	+
P2.3.3	Awarded professor titles to non-employees count	5	+	+	+	+
P2.3.4	Scientific advisory to non-employees count	1	+	+	+	+
P2.4	**Other achievements indicating potential**		+	+	+	+
P2.4.1	Membership in the governing bodies of international scientific organizations count	1,2	+	+	+	+
P2.4.2	Editor-in-chief of journal indexed on the JCR or the ERIH count	2	+	+	+	+
P2.4.3	Editor of journal indexed on the JCR or the ERIH count	1	+	+	+	+
P2.4.4	Membership in expert panels count	2	+	+	+	+
P2.4.5	Publishing a journal indexed on the JCR or the ERIH count	3	+	+	+	+
P2.5	**R&D projects volume**	*	-	–	+	-
P2.6	**Laboratories count**		-	+	+	-
P2.6.1	National certification and accreditation of laboratories count	10	-	+	+	-
P2.6.2	International certification and accreditation of laboratories count	10	-	+	+	-
P2.7	**Status of National Research Institute**	**10**	-	+	+	-

Table 1 (*Continued*)

Metric 3: Material effects of the scientific activity (M3)

ID	Parameter Name	Points	SSH	SE	LS	ASP
P3.1	**National and international R&D projects volume**		+	-	-	+
P3.1.1	International R&D projects volume	1***	+	-	-	+
P3.1.2	National R&D projects volume	0.5***	+	-	-	+
P3.2	**Financial incomes volume**		-	+	+	-
P3.2.1	Salaries financed by R&D projects volume	2***	-	+	+	-
P3.2.2	Purchase or development of scientific equipment volume	2***	-	+	+	-
P3.3	**Sales of research results volume**		**+**	+	-	+
P3.3.1	New results ordered by third-party volume	1***	-	+	-	+
P3.3.2	Sale of know-how licenses volume	1***	-	+	-	+
P3.3.3	Business consultancy volume	1***	**+**	+	-	+
P3.4	**Implementation of research results volume (only for applied research institutes)**	0.1***	-	+	+	-
P3.5	Contribution to national defence	****	-	+	-	-

Metric 4: Other effects of the scientific activity (M4)

ID	Parameter Name	Points	SSH	SE	LS	ASP
P4	Other effects of the scientific activity	**0–100**	+	+	+	+

Annotation: groups of sciences: SSH – social sciences and the humanities, SE – sciences and engineering, LS – life sciences and ASP – art sciences and artistic production. Mark (+) signifies that a parameter was included in the evaluation of a given group of science. The ranges of points were the same for each group of science.

* Applies only to architecture, urban planning and art design.

** Congress languages – English, German, French, Spanish, Russian, Italian or a fundamental language for a discipline, e.g. Czech for Czech philology.

*** For every 50,000 PLN.

**** Points were assigned by a special evaluation group.

* max 200 points related to a percentage of incomes to the government dotation.

Sciences from the SSH group were assigned to their single JEG. In all, 60 JEGs were established, and the evaluation results of the scientific units were compared within a particular JEG. The number of scientific units within a particular JEG ranged from one to 93. The number of scientific units within a JEG depended not only on the science group and the type of scientific unit but also on the field of science in terms of the government classification of scientific disciplines in Poland.

The evaluation was performed in two phases. In the first phase, the research outcomes were assessed by parametric evaluation according to Metrics 1–3 and by expert evaluation within Metric 4. The results of the first phase were assigned to one of three scientific categories in each scientific unit, as follows:

- Category A – Very good level
- Category B – Acceptable level with the recommendation to strengthen the scientific activity
- Category C – Unsatisfactory level

Categories A, B and C were assigned from the result of pairwise comparisons within a JEG, more details in Koczkodaj et al. (2014).

In the second phase, an A+ category (the leading level in the country) was assigned to the best units from Category A in the first phase, following additional metrics-informed expert judgment. The experts decided which set of metrics to use for a certain group of scientific units. Usually, they were using citation count, h-index and publication in top journals. The assigned category plays a major role in the distribution of funds for science. As a result, 37 scientific units obtained the highest category, that of A+; 308 units obtained Category A; 541 units obtained Category B and 77 Category C.

3. Methods

3.1. Data

We examined the research question using aggregated data from the Polish Ministry of Science and Higher Education. During the last evaluation, all scientific units had to submit the *Scientific Unit Questionnaire* through the *POL-on – Information System on Higher Education*. In this questionnaire, a scientific unit had to assign their research outcomes – that is, their evaluation items – to the parameters (see Table 1). The submitted data and assignment to the parameters were the evaluation basis for the experts. Experts could question some evaluation items if they did not meet various formal metrics. All evaluated data (evaluated items) were translated into points according to the metrics and parameters.

The data were aggregated at the scientific unit level as well as at the Joint Evaluation Group level. Note that we do not have full access to data for individual researchers and their research activity. Moreover, we do not have full data of the complete productivity of a given scientific unit: we have data about the evaluation items that were included in the process

of evaluation. The final dataset consists of the number of points assigned to each parameter, for example, P2.4.5: *Publishing in a journal indexed on the JCR or the ERIH*, for a given scientific unit. Each scientific unit is assigned to only one Joint Evaluation Group.

In our analysis, we used the results of the evaluation from September 2013, meaning the same data, which were used during the 2013 evaluation process. Only the results of the main metrics (M1–M4) were publicly presented.

The analysed dataset consists of information about 962 scientific units participating in the 2013 evaluation. One scientific unit was omitted, the Polish Academy of Learning in Cracow, because it followed a dedicated evaluation procedure. The dataset consists of a table with points assigned to every scientific unit for all four metrics and 65 parameters (see Table 1). In the analyses, we have assumed that we are at the first phase of evaluation – in other words, we have to assign scientific units to one of three categories: A, B or C. We do not take into account the other phase in which an expert-based evaluation was carried out for distinguishing the A+ category.

In the analysis, we have used the *primary parameters*, namely the parameters with data directly submitted by the scientific units, such as P1.1.1 (Article in a journal indexed in the Journal Citation Reports). In the dataset, this parameter contains information about the sum of points that a scientific unit has obtained for articles in journals indexed in the JCR.

The *aggregated parameters* are those that were not supplied by a research unit but were a summation of the primary or lower level (more detailed) parameters like P1.1 (Journal articles), which is a sum of P1.1.1, P1.1.2, P1.1.3, P1.1.4 and P1.1.5.

Finally, the metric is again the sum of the aggregated and the primary parameters. For example, Metric 1 is a sum of P1.1, P1.2, P1.3, P1.4 and P1.O. Moreover, according to the regulations of evaluation, Metrics 1 and 3 are divided by N (the four-year average full-time employment equivalent).

In Metric 1, the number of evaluation items was limited to $3N - 2N_0$ (see section 2.2). A scientific unit may have reported as many evaluation items as possible, but only a limited number of them were taken into account for the evaluation and played a role in the final result. Within the limit, the evaluation items with the highest number of points were included (compare Table 1). This means that when a scientific unit has been able to provide many articles from journals indexed in the JCR (10–50 points), monographs (20–25 points) or patents (25 points), there were fewer spots for those eventual evaluation items with a lower number of points, such as chapters in monographs (4 points) or patent applications (2 points). Moreover, the percentage of included monographs was limited to 40% of all submitted publications for evaluation by a scientific unit from the SSH and to 10% for scientific units from the other three groups of sciences. The number of possible points that a scientific unit could obtain in parameter P2.4 (Other achievements indicating potential) was limited to 50. The rest of the parameters in M1–M4 were unrestricted; in other words, a scientific unit could obtain as many points as possible for its submitted research outcomes.

3.2. Data analysis

All statistical analyses were performed in R software using the following packages: MASS (Venables & Ripley, 2002), rms (Harrell, 2014), rpart (Therneau, Atkinson, & Ripley, 2015) and caret (Kuhn, 2015). Our main goal was an analysis of the relations between the parameters and an analysis of the influence of particular parameters in the assigned scientific categories to the scientific units. We computed Pearson correlation coefficients between 68 metrics and parameters to identify the relationships within the dataset.

Subsequently, we focused on the selection of those variables that were significant for the evaluation process of scientific units. Categorization of the scientific units can be perceived as a problem of classification in which a decision is understood as a selection into one of the three categories: A, B or C. To examine whether there are correlations between the various metrics and the parameters within a given group of science, we used two models: multivariate logistic regression and a decision tree model. These models have a simple structure and can provide a clear interpretation. The quality of the prediction was not our priority. One can improve the prediction by using more complicated models like SVM, neural networks or ensemble methods. However, such “black-box” models are very difficult to interpret.

The scientific category was used as the output variable, and all primary parameters were used as predictors. Primary parameters from M1 and M3 were divided by N (arithmetic mean of the number of academic staff members) to remain coherent with the evaluation regulations.

Two multivariate logistic regression models were built for each scientific group. The first model recognized the A category versus the joint B and C category. The other model distinguished between the B category and the C category. Each multivariate logistic regression model was fitted, starting from the whole set of predictors (primary parameters) and from a set of predictors with excluded parameters belonging to M2. The idea of excluding all parameters from M2 is a result of the Pearson correlation coefficient analysis. Next, predictors were reduced by a stepwise backward procedure using AIC metric. Finally, the estimation of the accuracy was performed with a 10-fold cross-validation procedure. We were able to use ridge regression with a small value of penalty parameter thanks to the MASS and rms packages (Harrell 2014; Venables & Ripley, 2002) – because of the relatively large number of potential predictors with respect to the sample size within each group of science.

Decision tree models were fitted starting from the whole set of predictors. The size of the decision tree (the complexity parameter) was chosen via a cross validation procedure (thanks to the rpart and caret packages). The information gained was used as a split metric.

Table 2
The highest Pearson correlation coefficients between metrics and parameters.

Metric/Parameter		Metric/Parameter		Pearson correlation coefficient	95% confidence interval
P3.3.1	New results ordered by third parties	P3.4	Implementation of research results	0.861	(0.791; 0.908)
P1.2.1	Monograph in a congress language	P1.2.5	Edited volume in Polish	0.777	(0.750; 0.801)
M1*	Scientific and creative achievements	M2	Scientific potential	0.774	(0.748; 0.798)
P2.2	Scientific development of employees	P2.3	Academic promotions	0.749	(0.720; 0.775)
P2.2	Scientific development of employees	P2.3.3	Professor titles of non-employees	0.737	(0.706; 0.764)
P2.2.2	DSc degrees of employees	P2.3.3	Professor titles of non-employees	0.73	(0.699; 0.758)
P1.1	Journal articles	P2.2	Scientific development of employees	0.727	(0.696; 0.756)
P2.2.2	DSc degrees of employees	P2.3	Academic promotions	0.726	(0.695; 0.755)
P1.1	Journal articles	M2	Scientific potential	0.713	(0.681; 0.743)
P2.2.1	PhD degrees of employees	P2.2.2	DSc degrees of employees	0.714	(0.682; 0.744)
P2.6.1	National certification and accreditation of laboratories	P2.6.2	International certification and accreditation of laboratories	0.658	(0.608; 0.703)
P1.1.1	Article in a journal indexed in the Journal Citation Reports (JCR)	M2	Scientific potential	0.616	(0.576; 0.654)

* M1 was not divided by N.

Table 3
Pearson correlation coefficients between M1 and M2 by groups of sciences.

Group of science	Pearson correlation coefficient	95% confidence interval	Sample size
Social sciences and humanities (SSH)	0.894	(0.869; 0.915)	303
Sciences and engineering (SE)	0.764	(0.714; 0.806)	233
Life sciences (LS)	0.865	(0.828; 0.894)	323
Art sciences and artistic production (ASP)	0.399	(0.223; 0.550)	103

4. Results

In this section, we first address the correlations between the metrics and the parameters of the evaluation, followed by multivariate regression models. Then we present the decision trees and the effects of the Metric 2 removal.

4.1. Correlation analysis

In Table 2 we have presented the 12 correlations with the highest Pearson coefficients from the whole dataset. There were 40 pairs of metrics or parameters with a correlation coefficient higher than 0.6. Moreover, 25 of these 40 highest correlations were between parameters within M2 itself, eight were between parameters within M1 and M2, three were between parameters within M1 itself, three were between parameters within M3 itself and one was between parameters in M1 and M3.

One of the highest correlation coefficients was between M1 (Scientific and creative achievements) and M2 (Scientific potential), the coefficient was 0.774, with a 95% confidence interval (0.748; 0.798). Other parameters within M1 were also correlated with M2. For instance, the correlation coefficient for parameter P1.1 (Journal articles) was 0.713, while the 95% confidence interval was (0.681; 0.743).

Let us have a closer look at the relationship between the metrics M1 and M2. Table 3 presents the Pearson correlation coefficients broken into the four groups of sciences. The correlation coefficient between M1 and M2 for the whole dataset was 0.774 and for three out of the four scientific groups even higher: for social sciences and humanities 0.894, for sciences and engineering 0.764 and for life sciences 0.865. For art sciences and artistic production, it was 0.399.

4.2. Multivariate logistic regression models

Table 4 shows the multivariate logistic regression models. It is noteworthy that they have quite a good accuracy, over 80%. We have presented R2 and AIC measures to show how the models were changed after the exclusion of the parameters from M2. In general, the accuracy of prediction after this procedure was quite similar. However, one should note that the stepwise procedure in the case of the models with all parameters sometimes also leaves as relevant also parameters from

Table 4
Multivariate logistic regression models.

Parameter	SSH				SE				LS				ASP				Times used
	A vs. BC		B vs. C		A vs. BC		B vs. C		A vs. BC		B vs. C		A vs. BC		B vs. C		
	All	All-M2	All	All-M2	All	All-M2	All	All-M2	All	All-M2	All	All-M2	All	All-M2	All	All-M2	
P1.1.1	***	***	+	*	***	***	**	***	***	***	***	***		+	+	+	15
P1.1.2	***	***	***	***	*	***		***	+	+	+		**	+	+		13
P1.1.3	***	***			+			+		+		+	+	+		+	9
P1.1.4	–	–	–	–	**	***	+	+		+			–	–	–	–	5
P1.1.5	***				–	–	–	–	–	–	–	–	**				2
P1.2.1	***	***		**	+	*		+					+	+			8
P1.2.2	+	***		***			***	**	+		+			+			8
P1.2.3					+						+		*	+			4
P1.2.4			*					*	+			+		+			5
P1.2.5	+	+	+						*	+	+	+	**	+			9
P1.2.6	+		**	+			+			+			+	+			7
P1.3.1	–	–	–	–	**	***	+	*	**	*	+						7
P1.3.2	–	–	–	–													0
P1.3.3	–	–	–	–				+	+			+		+			4
P1.3.4	–	–	–	–		**		+							+		3
P1.3.5	–	–	–	–			+		+	+		+					4
P1.4.1	–	–	–	–				*	–	–	–	–		*	+	+	4
P1.4.2	–	–	–	–	+	+			–	–	–	–				**	3
P1.4.3	–	–	–	–	**				–	–	–	–		***	+	*	4
P1.4.4	–	–	–	–					–	–	–	–				*	1
P1.4.5	–	–	–	–					–	–	–	–			+	+	2
P1.4.6	–	–	–	–	+	+	***		–	–	–	–					3
P1.4.7	–	–	–	–	*	*			–	–	–	–					2
P1.4.8	–	–	–	–					–	–	–	–					0
P1.O	–	–	–	–	*	*			–	–	–	–					2
P2.1.1									+				*				2
P2.1.2	+						+		+				+				4
P2.2.1									**				+				2
P2.2.2	+		**						+		+		+				5
P2.2.3	*																1
P2.3.1											+		*				2
P2.3.2					*		**		+				+		+		5
P2.3.3											+				+		2
P2.3.4			+		*								+				3
P2.4.1	+						+		+		+		+		+		6
P2.4.2	+				*		**				*						4
P2.4.3									+								1

P2.4.4	+								+				+				3
P2.4.5	***																1
P2.5	–	–	–	–	–	–	–	–	+				–	–	–	–	1
P2.6.1	–	–	–	–					+		+		–	–	–	–	2
P2.6.2	–	–	–	–									–	–	–	–	0
P2.7	–	–	–	–					*				–	–	–	–	1
P3.1.1	+	***			–	–	–	–	–	–	–	–					2
P3.1.2	+				–	–	–	–	–	–	–	–	+				2
P3.2.1	–	–	–	–			**	+	+	*		+	–	–	–	–	5
P3.2.2	–	–	–	–	**	*			+	*	+		–	–	–	–	5
P3.3.1	–	–	–	–	**	***		*	–	–	–	–					3
P3.3.2	–	–	–	–					–	–	–	–					0
P3.3.3	+		**		*	*			–	–	–	–					4
P3.4	–	–	–	–	+	+			*			+	–	–	–	–	4
P3.5	–	–	–	–				**	–	–	–	–	–	–	–	–	1
P4	+	+	+		+	+	+	+	***	***	***	***	+	*	+	**	15
Observations	303	303	221	221	323	323	207	207	233	233	145	145	103	103	82	82	
Accuracy	0.931	0.921	0.946	0.982	0.817	0.811	0.932	0.937	0.888	0.88	0.910	0.862	0.825	0.825	0.841	0.951	
AUC	0.997	0.987	0.999	1.000	0.942	0.932	0.997	0.996	0.993	0.972	0.983	0.958	0.985	0.971	0.990	0.996	
R2	0.914	0.855	0.842	0.873	0.711	0.673	0.854	0.853	0.89	0.801	0.739	0.647	0.812	0.757	0.658	0.669	
AIC	40.01	100.82	20	12	231.49	241.05	28.03	32	91.67	125.35	30	60.93	38	53.91	22	18	
Number of variables	28	14	28	14	49	32	49	32	38	20	38	20	44	30	44	30	
Number of significant variables	19	8	9	5	20	15	13	15	23	11	14	9	18	13	10	8	
Number of insignificant variables	9	6	19	9	29	17	36	17	15	9	24	11	26	17	34	22	
Relevant to all variables ratio	68%	57%	32%	36%	41%	47%	27%	47%	61%	55%	37%	45%	41%	43%	23%	27%	

Annotation: In the columns labelled "−M2", there are models without parameters from Metric 2. Observations: the number of observations within a group of science (including all types of scientific units); Accuracy: the accuracy of the prediction, i.e. the probability that the model recognizes the category of unit correctly, estimated via a 10-fold cross-validation procedure; AUC (Area Under [ROC] Curve); pseudo R2 is the analogue or $R\hat{2}$ metric for the logistic regression (defined as: 1-L(M)/L(1), where L(M) is the likelihood of fitting the full model and L(1) is the likelihood of the model fitting only to the intercept); AIC is Akaike Information Criterion (AIC = 2 k − 2 ln(L(M))), where k is the number of model parameters (Harrell, 2014; Hosmer, Lemeshow, & Sturdivant, 2013). The variables which appear in the model are marked by "+", additionally statistically significant variables are marked by asterisks (*p-value less than 0.05, **p-value less than 0.01 and ***p-value less than 0.001), variables not included are marked with "–".

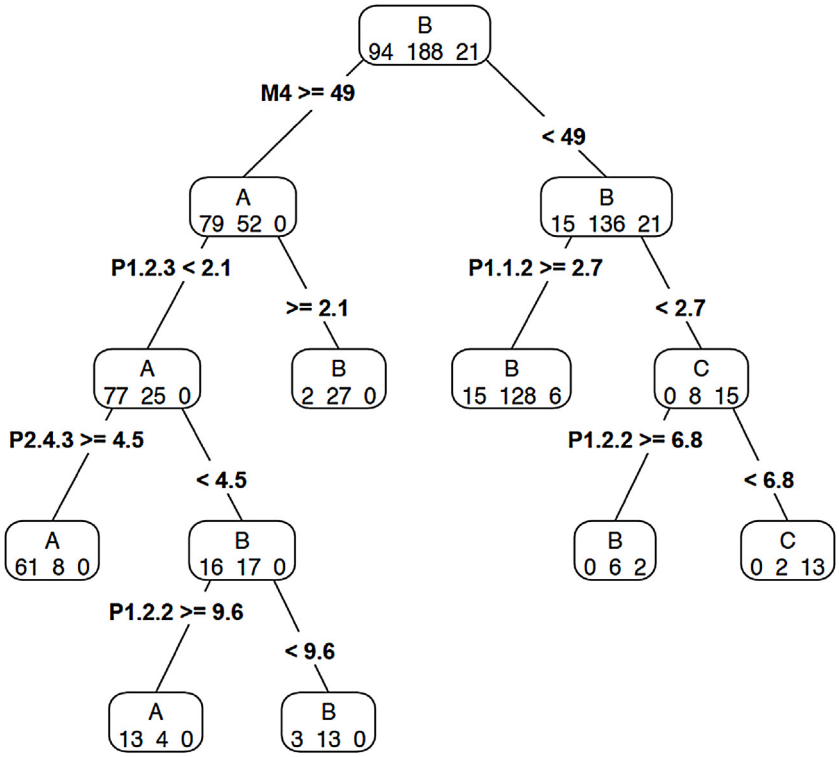

Fig. 1. Decision tree for the SSH (social sciences and humanities) group.

the M2. This could be interpreted in two ways. These parameters that remain in the model could, in reality, be statistically significant, or they could be in correlation with the excluded parameters from M1.

In the evaluation, from 28 to 49 primary parameters were used depending on the group of sciences. The number of parameters results from the Regulation of 13 July 2012 on the metrics and procedure for assigning scientific categories to scientific units (see Table 1). The stepwise selection with the multivariate logistic regression models revealed that only some of the primary parameters are actually statistically significant. In SE, for instance, only 20 of the 49 primary parameters were relevant for separation into Category A, and only 13 were relevant for separation into Category B from Category C. In general, there were fewer variables needed to distinguish Category C than to distinguish Category A.

In the models without parameters from M2, fewer variables were relevant than in the model with all parameters. However, distinguishing Category C from B in SE required 13 variables, and the model A vs. B and C required 15 parameters. At the same time, the prediction accuracy stayed almost untouched or even improved in the case of the separation of the C category models (except for the LS).

Among the 53 primary parameters, only 10 were common for all four groups of sciences. Among these 10 parameters, P1.1.1 and M4 were significant in 15 models, P1.1.2 in 13 models and P1.1.3 and P1.25 were significant in 9 models. At the other end of the scale, 22 variables performed at most in two models.

It is interesting that the parameters from P1.4 were significant for scientific units from SE, even though only a very limited number of 26 among 323 scientific units (which do research in architecture, urban planning or art design) could report their outputs there.

4.3. Decision trees

Figs. 1–4 present the decision trees for the four groups of sciences. In the frames are the dominant scientific category and the number of scientific units grouped by the category (A, B or C) in the lower line. On the branch are the given metrics or parameters that split the higher-level set of scientific units into the two lower-level subsets. The accuracies of the tree models are respectively 0.78 for SSH, 0.65 for SE, 0.77 for LS and 0.66 for ASP. The present accuracies are lower than the above-mentioned accuracies of the multivariate logistic regression models. However, the decision trees give us insight into which variables are the most important from the perspective of the assigned categories. The accuracy of the decision trees could be improved by adding more specific decision rules, which would provide further tree leaves. The significant variables chosen by the multivariate logistic regression models also appear in the decision trees.

Let us analyse the SSH decision tree (Fig. 1). In this group of science, there are 94 scientific units assigned to the A category, 188 in the B category and 21 in the C category. The first decision rule was M4: if it was higher than or equal to 49, then a given

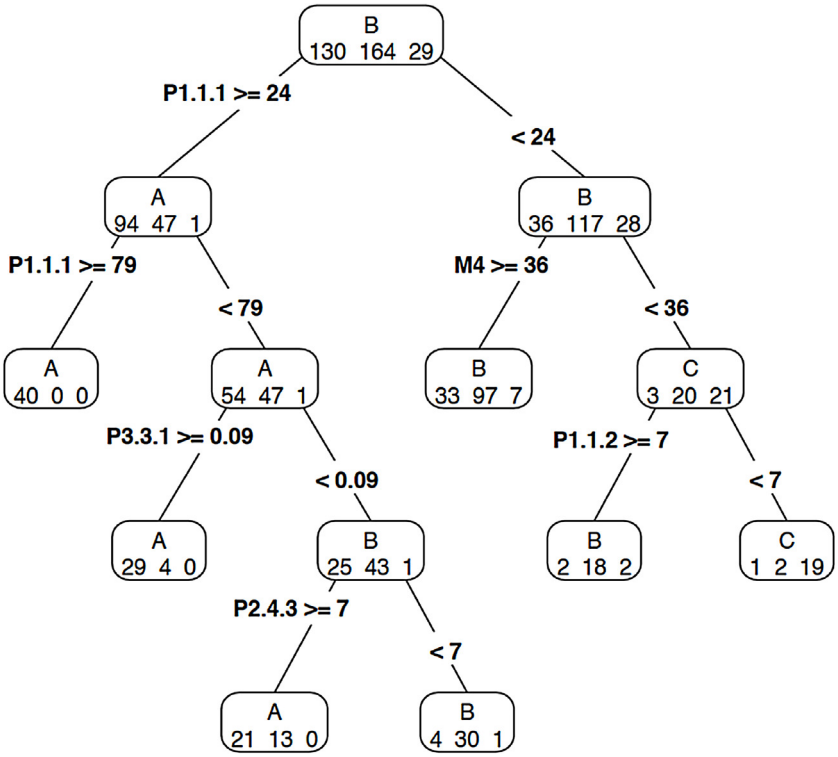

Fig. 2. Decision tree for the SE (sciences and engineering) group.

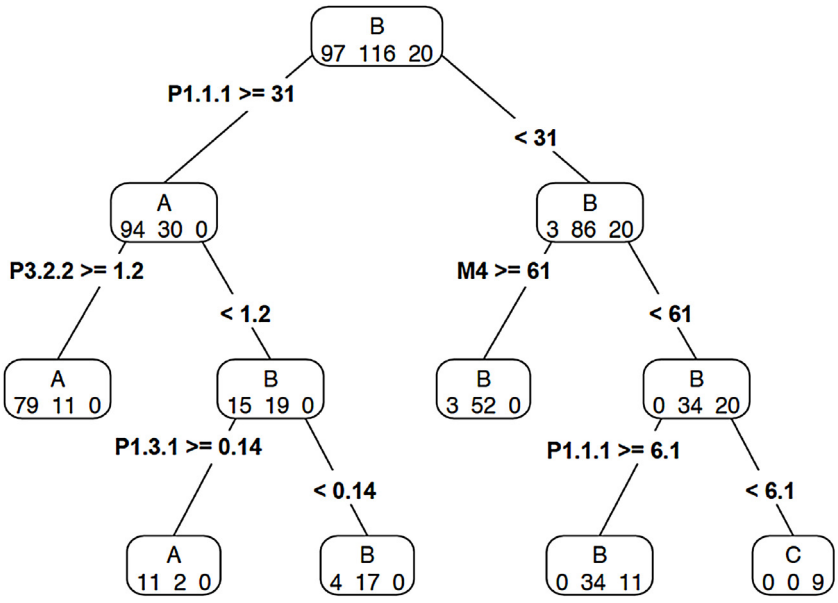

Fig. 3. Decision tree for the LS (life sciences) group.

scientific unit followed the left-hand branch, resulting in a set of scientific units dominant in the A category. Otherwise, it followed the right-hand branch, resulting in a set of scientific units dominant in the B category. The second decision rule for the left-hand set is P1.2.3: if it was lower than 2.1, then a given scientific unit followed the left-hand branch, or, if higher than or equal to 2.1, it followed the right-hand branch. In the whole tree, we had six decision rules and seven leaves – in other words, no further sets. The most left-handed leaf contains 69 scientific units among which 61 were in the A category and eight in the B category. The A category was dominant. In this set, the eight scientific units assigned to the B category are probably very similar to the A category units. Thus, the next specific decision rules might have separated them. It is worth

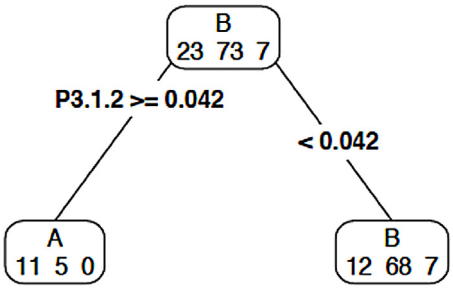

Fig. 4. Decision tree for the ASP (art sciences and artistic production) group.

Table 5
Categories without Metric 2 by groups of sciences.

Categories		SSH	SE	LS	ASP
A	with Metric 2	94	130	97	23
	without Metric 2	74	141	98	19
B	with Metric 2	188	164	116	73
	without Metric 2	203	158	116	69
C	with Metric 2	21	29	20	7
	without Metric 2	26	24	19	15

noting that in the decision trees, we did not take into account the JEGs, which are more homogenous. The most important variables in SSH were M4 and the parameters from M1. The parameters from M2–P2.4.3–appeared only once.

In the SE group, the most important decision rule was P1.1.1 higher than or equal to 24 points (Fig. 2). The other most important rule was again P1.1.1. However, this time the parameter in question had to be higher than or equal to 79. In the SE tree, there were six decision rules from all four metrics. Nonetheless, a parameter from M2 was on the fourth level of the tree. Despite the strong heterogeneous character of the scientific units from this group of sciences, the accuracy for this decision tree was 0.65. Some sets of units are hardly separable (like the set with 33 scientific units in the A category, 97 in the B category and 7 in the C category). Scientific units on the level of the JEGs are much more homogeneous, but, in the majority of cases, the JEG size was not large enough to build a reasonable decision tree.

There were five decision rules from M1, M2 and M3 in the LS tree (Fig. 3). As in the SE tree, parameter P1.1.1 was the most important, but the cut-off point was at a higher level (31 points). Next in importance were decision rules P3.2.2 and M4.

In our analysis, we had the most problems with the decision tree for the ASP group (Fig. 4). There was only one decision rule P3.1.2 higher than or equal to 0.042, which did not give a good separation, and expansion of the tree via additional decision rules did not clear the problem up. Nevertheless, it is noteworthy that in the ASP group, the most important parameter was related to the money that a given scientific unit was able to acquire for grants in competitive procedures, usually with the expert-based proposals rating.

M4 appears in three decision trees (SSH, SE and LS). In the SSH tree, it is the most important rule that divides scientific units into smaller sets of very good-level units (79 in the A category and 52 in the B category) from the rest (mostly B and C categories). In the case of the SE and LS trees, this metric did not play as significant a role and allowed separating the acceptable-level units (the B category) from the unacceptable-level units (the C category).

The results of the correlation analysis (Tables 2 and 3), of the multivariate logistic regression models (Table 4) and of the decision trees (Figs. 1–4) provide arguments for answering our main research question: how much data does a performance-based research funding system need? In the Polish system, not all the collected data are necessary to achieve the main goal of the system, which is a categorization of scientific units. As the results show, the parameters within M2 (Scientific potential) play a marginal role in the decision trees (see Figs. 1–4) but also M2 itself is highly correlated to M1. Other metrics (M1, M3, M4) play an important role in distinguishing the performance levels of the Polish scientific units.

4.4. Effects of the Metric 2 removal

We have simulated an effect of the Metric 2 (Scientific potential) removal from the Polish PRFS. To assign new scientific categories – that is, without using M2–we have recalculated the pairwise comparisons within a given JEG in the same way it was done in 2013. By removing M2 from the system, the assigned scientific category in 2013 would change only for 79 scientific units, among which 49 would receive a lower category and 30 a higher category. In the SSH, 29 units would receive lower and four higher categories; in the SE, four would receive lower and 20 higher categories; the LS numbers are 1 and 3 and finally for the ASP 15 and 3, respectively. The detailed results are presented in Table 5.

Removing Metric 2 from the system would concern two types of scientific units. The first type are those which have moderate or even inferior results in Metric 1 in comparison with other scientific units belonging to the same category (in terms of the categories assigned in 2013) and at the same time very well performing in Metric 2. They would be assigned to a lower scientific category. The other scientific units are those, which would be assigned to a higher scientific category. They are performing above the average in Metric 1 and below the average in Metric 2.

5. Discussion

The Polish PRFS is perceived by the Polish scientific community as an overly complicated, formal and demandingly excessive effort in regard to supplying data. In fact, it is more data demanding than the PRFSs in Denmark, Finland, Norway and the UK. The Italian and Czech systems also require a large volume of data.

In our opinion, dismissing Metric 2 would help scientific units to focus on a smaller number of metrics. Our simulation has revealed that this would reinforce scientific units, which are good at publishing (Metric 1) and cooperation with the economic and scientific environment (Metric 3).

A large-scale statistical analysis is not an ample enough argument for deciding that some metrics can be left out. Our analysis can inform a decision, but the statistic tests cannot ultimately decide whether Metric 2 should be excluded from the Polish PRFS. There are no simple answers to the question of how to improve such a complex system. However, statistical implications can be used in the decision-making process on the political level. Poland needs more prestigious publications as well as more international cooperation because the Polish research outcomes are below the average EU levels in almost all rankings of higher education and science areas. Thus, solutions for prioritizing the activity of researchers and scientific units are needed.

On the overall level, excluding Metric 2 from the Polish PRFS may have various pros and cons. Some scientific units could start to prioritize the parameters included in Metrics 1 and 3 and at the same time abandon activities that are measured by Metric 2. Such a reduced system might leave less room for gaming, but, on the other hand, gaming might become more attractive in a simpler system. However, as Campbell (1979) wrote, "the more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor" (p. 85). In light of this quote, we can see the complexity in the decision on how to modify and improve a PRFS. If we assume that 'what gets measured gets done' and that these measures might be corrupted, then we should be also aware of the importance of a balanced discussion in regard to this challenge. Thus, our analysis and results might provide relevant arguments for such a discussion.

The other point we have drawn from our analysis is related to the number of outcomes about which the PRFS should collect information. The multivariate logistic regression models and the decision trees have shown that the low-impact outcomes were not important for the A category separation nor for the C category separation. The low impact outcomes are those for which a scientific unit is awarded a low number of points, such as from third and fourth quartile journals from the national scientific journal ranking, chapters, and patent applications.

Excluding low-impact outcomes would lead the Polish PRFS toward being an excellence-based research funding system that would be restricted to collecting only significant outcomes, especially in a metric related to scientific and creative achievements (M1 in the current system). If a scientific unit did not have enough good quality outcomes to fill the allocated slots (e.g. for publication), then the slots remain empty. This can reduce the number of evaluation items in the PRFS by 50%. Additionally, the proposed enhancements will facilitate the implementation of a science policy and will help scientists to focus on excellent research by clearly indicating what is supported by the state.

Our analysis, especially the multivariate models (see Table 4) and decision trees (see Figs. 1–4), shows that for two groups of science – the SE and the LS – there are similar significant variables and that both of these groups might be evaluated using the same set of metrics. The SSH and the ASP should be assessed with a different method. When we look closer at some disciplines in the SSH, such as psychology, it seems that they follow the publication patterns of SE and LS and may be evaluated according to those metrics. Moreover, the Polish system includes the ASP, in which the majority of outcomes are not publications. Thus, it might be worth considering reconfiguring the science groups and dividing SSH into two groups (humanities and social sciences), as in the OECD classification (2007). As a result, we would have a common metric set for the SE, the LS and social sciences. Additional analysis is needed concerning the most appropriate metrics for the humanities and the ASP.

In a perfect world, the head of a scientific unit, knowing the evaluation system, would be able to predict the financial consequences of actions that affect the scientific profile of his or her unit. It would be much easier to know what would happen when the quality of publications or the number of grants or citations improves. Furthermore, a simpler system leaves less room for gaming.

Over the last two decades, the research evaluation systems in search of measures of performance have been shifting away from analysing the *average values* of performance indicators and are instead looking at the *top values* of these indicators. As van Leeuwen, Visser, Moed, Nederhof, and Van Raan (2003) have shown, if we look for excellence in research, we should not use just one single indicator but rather a variety, and, what is more important, we should focus only on the most important outcomes among the other 'highly cited' publications and 'top articles'. Hicks (2009) has shown that the complexity of the research evaluation systems is caused by the regulatory processes (consultations) with stakeholders, such as decision-

makers, scientists and scientific managers, who try to keep a balance between the different groups of sciences. Focusing on only the most important parts of scientific research outcomes should be a major principle of the PRFSs.

The approaching 2017 evaluation in Poland for the period 2013–2016 will be carried out according to principles similar to those in 2013. The Ministry of Science and Higher Education in Poland and the Committee for Evaluation of Scientific Units have modified only some parameters. Larger changes concern Metric 2. New parameters have been added, such as coordination and participation in national and international research projects and the participation of employees in internships at leading scientific institutions. Not surprisingly, no parameters have been removed.

Future Polish excellence-based research funding systems could be further supplemented with parameters based on citations and international collaboration. The scientific units could garner both these parameters, but at the cost of additional effort. As a result, there is the risk that scientific units will pay too much attention to success in the evaluation system itself rather than on good research. Additionally, self-citation groups (Bonzi & Snyder, 1991; Teodorescu & Andrei, 2013) and fake collaborations (Wagner, 2005) are the noticeable challenges.

Even if we could implement the above-mentioned solutions, the important challenge caused by the unsatisfactory degree of coverage of non-English publications in the major bibliometric database still remains: how to evaluate and appreciate the humanities and arts as fields of science that have an important social impact.

We continue to search for the "Holy Grail": a system that is resistant to unanticipated distortions and that at the same time does not encourage "gaming" in the metrics.

6. Study limitations

Firstly, our study uses data submitted by the scientific units that are evaluated. The data are aggregated at the scientific unit levels, and their performance is translated into points according to the various metrics and parameters. We have used the results of our analysis to suggest which metrics or parameters may be excluded from the system. However, it should be highlighted that the importance of the metrics and parameters are determined by the range of points that are assigned to them, meaning that a performance which is assigned a lower number of points – for example, chapters – is only evaluated if a given scientific unit does not have enough good publications to submit and reach the limit $3N–2N_0$. Thus, defining the importance of metrics in terms of the assigned points is going to be one of the biggest challenges for the system as well as the issue that determines the future improvement of the system.

Second, our study relies on the publication count that is used in the Polish model – that is, whole counting where publications are not counted two or more times in one scientific unit. This means that modifying the way in which publications are counted might change the results of the categorization of scientific units. However, more importantly, modifying the publication count would most likely provoke a change in the behaviour of Polish researchers and their publication practices (Bloch & Schneider, 2016).

Finally, our analysis is based on the unit of analysis, meaning a scientific unit in which not all employees have to contribute to the performance of the unit. Thus, in some scientific units, only a few researchers "generate" the whole performance of their unit. If we want to transform the Polish model to excellence-based, we should provide regulations whereby all researchers from a given scientific unit have to submit their best outcomes. Only then can we measure the excellence of the whole scientific unit and not just the excellence of its best researchers.

7. Conclusions

With respect to the aim of the present paper and the research question, the following considerations can be derived from our analyses. The biggest challenge is how to improve the Polish model to achieve a two-fold aim: funding distribution to the best performers and motivating scientific units to increase their "excellence" in science. It is easy to redistribute block grants using a systematic approach. Nonetheless, increasing the motivation of scientific units is much more difficult. Another important question lies in how much of the funding should be distributed to the best performers and how much should remain to motivate those lesser successful scientific units.

The Polish PRFS has been criticized for its complexity. It comprises four metrics with 65 parameters. In total, 962 units reported information about one million research outcomes for the period 2009–2012; however, our analyses revealed that many of the parameters taken into account in the evaluation are closely correlated (especially within Metric 2) and that many of the parameters are not significant in regard to the category assignment.

The next cycle of evaluations will take place in 2017 and will be based on the regulations announced in 2016. For the upcoming evaluation for the period 2017–2020 that will take place in 2021, we propose reducing the number of metrics to three, primarily by removing Metric 2 (Scientific potential) from the system and limiting the number of journals on the national scientific journal ranking.

According to our analyses, we suggest a transition from a system in which scientific units report all their achievements to a system in which they show only the most important performances that meet the requirements of excellence in research. We are aware that some parameters might be included in a PRFS even if from a technical point of view it is not supported by data. Such a decision could allow for achieving broad support in the scientific community. Nonetheless, our aim is to suggest such an improvement of the Polish PRFS that could serve for goals established in the strategy for science policy (MNiSW, 2015).

At the same time, the policymakers should decide to invite the Polish academic community to participate in defining and measuring excellence. We must highlight here that the current and previous versions of the Polish PRFS used consultations with researchers; however, those consultations concerned mostly the scope of the metrics and the number of points assigned to the parameters. Finally, the result of those consultations was an exaggerated set of measures, which affects scientific units such that they report even their lowest performances to fulfil the allowed limits (e.g. the number of publications which could be submitted for evaluation).

Author contributions

EK undertook the literature search, co-drafted the paper and contributed intellectually to the development of the final manuscript. MK performed the data analysis and participated in the interpretation of the results and in manuscript preparation. PK conceived the original idea for the analysis, co-drafted the paper, contributed intellectually to the development of the final manuscript and contributed to the data analysis.

Funding

The work of EK was supported by the National Programme for the Development of Humanities in Poland [http://nprh.org]. Grant number 0057/NPHR3/H11/82/2014.

Acknowledgement

We would like to thank the Ministry of Science and Higher Education for its support in making the data available for our analyses.

References

Aagaard, K., Bloch, C., & Schneider, J. W. (2015). Impacts of performance-based research funding systems: the case of the Norwegian Publication Indicator. *Research Evaluation*, *24*(2), 106–117. http://dx.doi.org/10.1093/reseval/rvv003

Aagaard, K. (2015). How incentives trickle down: Local use of a national bibliometric indicator system. *Research Evaluation*, *42*(5), 725–737. http://dx.doi.org/10.1093/scipol/scu087

Abramo, G., & D'Angelo, C. A. (2015). Evaluating university research: Same performance indicator, different rankings. *Journal of Informetrics*, *9*(3), 514–525. http://dx.doi.org/10.1016/j.joi.2015.04.002

Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., et al. (2015). Evaluating scientific research in Italy: The 2004–10 research evaluation exercise. *Research Evaluation*, 1–14. http://dx.doi.org/10.1093/reseval/rvv008

Antonowicz, D., & Brzeziński, J. M. (2013). Doświadczenia parametryzacji jednostek naukowych z obszaru nauk humanistycznych i społecznych 2013–z myślą o parametryzacji 2017. *Nauka*, *4*, 51–85.

Arthur, M. (2015). Excellence in research. In O. Tayeb, A. Zahed, & J. Ritzen (Eds.), *Becoming a world-Class university* (pp. 77–90). Cham: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-26380-9_5

Bloch, C., & Schneider, J. W. (2016). Performance-based funding models and researcher behavior: An analysis of the influence of the Norwegian Publication Indicator at the individual level. *Research Evaluation*, 047–112. http://dx.doi.org/10.1093/reseval/rvv047

Bonzi, S., & Snyder, H. W. (1991). Motivations for citation: A comparison of self citation and citation to others. *Scientometrics*, *21*(2), 245–254. http://dx.doi.org/10.1007/bf02017571

Campbell, D. T. (1979). Assessing the impact of planned social change. *Evaluation and Program Planning*, *2*, 67–90.

Chowdhury, G., Koya, K., & Philipson, P. (2016). Measuring the impact of research: Lessons from the UK's research excellence framework. *PLoS One*, *2014*, 1–15. http://dx.doi.org/10.1371/journal.pone.0156978

Ferrara, A., & Bonaccorsi, A. (2016). How robust is journal rating in Humanities and Social Sciences? Evidence from a large-scale, multi-method exercise. *Research Evaluation*, rvv048-13. http://dx.doi.org/10.1093/reseval/rvv048

Good, B., Vermeulen, N., Tiefenthaler, B., & Arnold, E. (2015). Counting quality? The Czech performance-based research funding system. *Research Evaluation*, *24*(2), 91–105. http://dx.doi.org/10.1093/reseval/rvu035

Haddow, G., & Genoni, P. (2010). Citation analysis and peer ranking of Australian social science journals. *Scientometrics*, *85*(2), 471–487. http://dx.doi.org/10.1007/s11192-010-0198-4

Harrell, F. E., Jr. (2014). *rms: Regression modeling strategies*. http://CRAN.R-project.org/package=rms

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431. http://dx.doi.org/10.1038/520429a

Hicks, D. (2009). Evolving regimes of multi-university research evaluation. *Higher Education*, *57*(4), 393–404. http://dx.doi.org/10.1007/s10734-008-9154-0

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, *41*(2), 251–261. http://dx.doi.org/10.1016/j.respol.2011.09.007

Hosmer, D. W. J., Lemeshow, J., & Sturdivant, R. X. (2013). *Applied logistic regression*. Hoboken, NJ, USA: John Wiley & Sons. http://dx.doi.org/10.1002/9781118548387

Jonkers, K., & Zacharewicz, T. (2016). Research *Performance based funding systems*: *A comparative assessment*. 10.2791/659483.

Klincewicz, K. & Szkuta, K. (2016). *RIO Country Report 2015: Poland*; EUR 27872 EN; 10.2791/984739.

Koczkodaj, W. W., Kułakowski, K., & Ligęza, A. (2014). On the quality evaluation of scientific entities in Poland supported by consistency-driven pairwise comparisons method. *Scientometrics*, *99*(3), 911–926. http://dx.doi.org/10.1007/s11192-014-1258-y

Kuhn, M. (2015). Caret: Classification and regression training. *Astrophysics Source Code Library*, *1*, 05003.

Kulczycki, E., Drabek, A., & Rozkosz, E. A. (2015). Publikacje a zgłoszenia ewaluacyjne, czyli zniekształcony obraz nauki w Polsce. *Nauka*, 35–58.

Kulczycki, E. (2017). Assessing Publications through a Bibliometric Indicator: The case of comprehensive evaluation of scientific units in Poland. *Research Evaluation*, 1–12. http://dx.doi.org/10.1093/reseval/rvw023

MNiSW (Ministry of Science and Higher Education in Poland) (2015). *Higher Education and Science Development Programme for 2015–2030* (orig. Program Rozwoju Szkolnictwa Wyższego i Nauki na lata 2015–2013), Semptember 2015.

OECD (2007). Organisation for Economic Co-operation and Development. R*evised field of science and technology (FOS) classification in the Frascati manual*. Paris; 2007 Feb. Report No.: DSTI/EAS/STP/NESTI(2006)19/FINAL. Available from: https://www.oecd.org/science/inno/38235147.pdf.

Rijcke, S., de Wouters, P. F., Rushforth, A. D., Franssen, T. P., & Hammarfelt, B. (2016). Evaluation practices and effects of indicator use: A literature review. *Research Evaluation*, *25*(2), 161–169. http://dx.doi.org/10.1093/reseval/rvv038

Saarela, M., Kärkkäinen, T., Lahtonen, T., & Rossi, T. (2016). Expert-based versus citation-based ranking of scholarly and scientific publication channels. *Journal of Informetrics*, *10*(3), 693–718. http://dx.doi.org/10.1016/j.joi.2016.03.004

Sadowski, I., & Mach, B. W. (2014). Parametryzacja i kategoryzacja jednostek naukowych w roku 2013 jako praktyka ewaluacyjna i proces instytucjonalny. *Nauka*, *2*, 67–103.

Schneider, J. W., Aagaard, K., & Bloch, C. W. (2014). What happens when funding is linked to (differentiated) publication counts? New insights from an evaluation of the Norwegian Publication Indicator. In E. Noyons (Ed.), *Proceedings of the science and technology indicators conference 2014 Leiden Context Counts: Pathways to Master Big and Little Data* (pp. 543–550). Leiden: Universiteit Leiden.

Sivertsen, G. (2015). Data integration in Scandinavia. *Scientometrics*, 1–7. http://dx.doi.org/10.1007/s11192-015-1817-x

Skoczeń, B., Antonowicz, D., Brzeziński, P., Jackowski, S., Pilc, A., & Zabel, M. (2014). *Kategoryzacja jednostek naukowych po kampanii odwołaní*. pp. 7–8. Forum Akademickie. https://forumakademickie.pl/fa/2014/07-08/kategoryzacja-jednostek-naukowych-po-kampanii-odwolan/

Sunkel, C. (2015). Excellence and the new social contract for science: In search for scientific excellence in a changing environment. *EMBO Reports*, *16*(5), 553–556. http://dx.doi.org/10.15252/embr.201540328

Teodorescu, D., & Andrei, T. (2013). An examination of citation circles for social sciences journals in Eastern European countries. *Scientometrics*, *99*(2), 209–231. http://dx.doi.org/10.1007/s11192-013-1210-6

Therneau, T., Atkinson, B., & Ripley, B. (2015). rpart: Recursive partitioning and regression trees.

Vanclay, J. K. (2011). An evaluation of the Australian Research Council's journal ranking. *Journal of Informetrics*, *5*(2), 265–274. http://dx.doi.org/10.1016/j.joi.2010.12.001

van Leeuwen, T. N., Visser, M. S., Moed, H. F., Nederhof, T. J., & Van Raan, A. F. J. (2003). The Holy Grail of science policy: Exploring and combining bibliometric tools in search of scientific excellence. *Scientometrics*, *57*(2), 257–280. http://dx.doi.org/10.1023/A: 1024141819302

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S*. Springer.

Wagner, C. S. (2005). Six case studies of international collaboration in science. *Scientometrics*, *62*(1), 3–26. http://dx.doi.org/10.1007/s11192-005-0001-0