Discrimination discovery in scientific project evaluation: A case study[☆]

Andrea Romei, Salvatore Ruggieri and Franco Turini

Dipartimento di Informatica, Università di Pisa
Largo B. Pontecorvo 3, 56127 Pisa, Italy

Abstract

Discovering contexts of unfair decisions in a dataset of historical decision records is a non-trivial problem. It requires the design of ad-hoc methods and techniques of analysis, which have to comply with existing laws and with legal argumentations. While some data mining techniques have been adapted to the purpose, the state-of-the-art of research still needs both methodological refinements, the consolidation of a Knowledge Discovery in Databases (KDD) process, and, most of all, experimentation with real data. This paper contributes by presenting a case study on gender discrimination in a dataset of scientific research proposals, and by distilling from the case study a general discrimination discovery process. Gender bias in scientific research is a challenging problem, that has been tackled in the social sciences literature by means of statistical regression. However, this approach is limited to test an hypothesis of discrimination over the whole dataset under analysis. Our methodology couples data mining, for unveiling previously unknown contexts of possible discrimination, with statistical regression, for testing the significance of such contexts, thus obtaining the best of the two worlds.

Keywords: Discrimination discovery, Gender bias, Case study, Situation testing, Data mining, KDD process

1. Introduction

Discrimination refers to an unjustified distinction of individuals based on their membership, or perceived membership, in a certain group or category disregarding individual merits. Unfair behaviors have been observed in a number of settings, including credit, housing, insurance, personnel selection and worker wages. Civil rights laws prohibit discrimination against protected groups defined on the grounds of gender, race, age, nationality, marital status, personal opinion, and so on. One crucial problem from legal, economic and social point of view is *discrimination discovery*, that is defining methods capable of providing evidence of discriminatory behavior in activities such as the ones listed above. In the socio-economic field the problem has been addressed by analysing data with a statistical approach. The basic idea is to see, by means of regression analysis, whether sensitive features, like gender and race, are correlated with a less favorable treatment of individuals. In our opinion such an approach can highlight only macroscopic situations, while missing to dig out situations of deep discrimination in (small) subsets of the population, i.e., niches of individuals with a particular combination of characteristics. As an example, consider the case of loan applications to a bank. The discriminatory behavior of a single branch office against applicants from a local minority can readily be hidden in the much larger set of decisions of the whole bank. In a few words, the statistical approach tends to find a general model characterizing the whole population, whereas discrimination often arises in specific contexts. We maintain that a data mining approach, that is the search for particular patterns in the data, must be coupled with statistical validation of the patterns found as a thorough strategy for discovering (unexpected or unknown) contexts of discrimination.

Data mining approaches to discrimination discovery have recently gained momentum, but, in our opinion, they still need major advancements: first, experimentation with real data; second methodological refinements; and third, the consolidation of a KDD process of discrimination discovery. Solving these issues is essential for the acceptance of discrimination discovery methods based on data mining in practice. In this paper we contribute to the advancement of the state-of-the-art in all those aspects.

First, we describe a large experiment on a real case study concerning the challenging problem of discovering gender discrimination in the selection of scientific projects for funding. Data refers to an Italian call for research proposals with 3790 applications nationwide.

Second, we couple a recently developed discrimination discovery method (Luong et al., 2011), based on data mining, with statistical validation of its results, thus reconciling the statistical and the data mining methodologies. The data mining method is unsupervised in the sense that there are no examples of discriminatory or non-discriminatory situations to learn from. Rather, the method discovers sets of situations in which the comparison of the features

[☆] A preliminary version of the results of this paper appeared in Romei et al. (2012).

and the decision may suggest a possible discrimination according to the legal methodology of situation testing. Statistical regression analysis is then used to prove or disprove them as an hypothesis.

Third, the description of the steps followed in the case study provides us with the basis for distilling a general KDD process for discrimination discovery. The process abstracted is rather complex. It contains both automated and semi-automated steps, the possibility of iterating subprocesses, and the need of tuning the parameters of the analyses.

The paper is organized as follows. Section 2 offers a survey of the background material, including a multidisciplinary introduction to the gender bias in scientific research, a survey of data mining approaches for discrimination discovery, and details on the approach based on situation testing. Sections 3 presents the case study of an Italian national research funding call, the available data, and the data preparation steps. Sections 5 and 6 report the core of the experiment, consisting of first extracting classification models describing contexts of possible discrimination, and then of validating such contexts by means of logistic regression. Four interesting contexts of possible discrimination are discussed. Section 7 generalizes the phases of the case study to a generic discrimination discovery process. Finally, the conclusions summarize the contributions of the paper and discuss some future work.

2. Background

The problem of gender differences in research is of concern to all major funding institutions. The European Union (EU) regularly publishes a report on the status of gender funding in the member states (European Commission, 2009), and it promotes gender equality in scientific research.[1] EU legislation includes an explicit resolution on women and science (Council of the E.U., 1999), which notably preceded the resolutions on racial and employment equality. The National Science Foundation (NSF) in the United States (US) supports the development of systemic approaches to increase the representation and advancement of women in academic science, technology, engineering, and mathematics through the ADVANCE program.[2] Broader overviews of studies and findings on gender in (scientific and technological) research have been conducted by the European Commission (2012) and by UNESCO (2007). In the next subsection, we review the existing literature on gender bias in scientific research, which basically relies on statistical regression as the basic tool for data analysis. Then, we briefly review recent approaches that use data mining for discrimination discovery and prevention. Finally, a deeper introduction is reported on the data analysis technique of situation testing, and to its implementation as a data mining algorithm.

2.1. Gender bias in scientific research

Forms of gender discrimination may explain women's under-representation in academia, both past and present. The surveys by Bentley and Adamson (2003) and Ceci and Williams (2011) cover multi-disciplinary literature on gender differences in manuscript reviewing, grant funding, university admission, and hiring and promoting in research. We focus here on grant and fellowship funding.

The influential paper by Wennerås and Wold (1997) reports a study on post-doctoral fellowship applications to the Swedish Medical Research Council (MRC) in 1995. A total of 62 applications were submitted by men and 52 by women: 16 men were funded (25.8%) versus 4 women (7.7%). Applicant's sex and scientific competence are considered as independent variables in a linear regression model estimating the score assigned by the reviewers. Scientific competence as a control factor is measured in terms of the number of published journal articles, their citation count, and total impact of those journals.[3]. The regression shows that "a female applicant had to be 2.5 times more productive than the average male applicant to receive the same score". However, subsequent studies by several funding societies in Europe and North America fail to show evidence of sex bias in approval rates (Ceci and Williams, 2011). In fact, Sandström and Hällsten (2008) analysed data on applications to the MRC in 2004 and found a *reversed* gender bias, namely a small but significant effect in favor of funding women's grants compared to men with the same scientific competence score.

Let us recall here a few large scale studies. RAND (2005) investigates grant applications in the period 2001-2003 to the NSF, the National Institutes of Health (NIH), and the Department of Agriculture. No evidence of gender bias was generally found after controlling for age, academic degree, institution, grant type, institute, and application year. There were two exceptions, partly explainable by the lack of further control variables. First, women received only 63% of the amount of funding awarded to men by the NIH. Secondly, women who applied in 2001 were less likely than men to submit applications in the next two years. Similar findings as in the first exception are also reported by Larivière et al. (2011) with reference to 9074 professors at universities in Quebec (Canada). The lower amount of funding received by women is not necessarily evidence of discriminatory decisions. Wilson (2004) explains the lower amount of funding granted to women by their marginalization within the scientific community, by their segregation to lower rank positions, and by their smaller social networks – all of these factors affecting their chances of funding possibilities.

[1] http://www.yellowwindow.be/genderinresearch

[2] http://www.portal.advance.vt.edu

[3] The problem of measuring and analysing science is the subject of *scientometrics* Indicators of scientific productivity of researchers have been debated for long time. See Bornmann and Daniel (2009) for a discussion on H-index and its variants, and for a comparison with the Impact Factor index.

Ley and Hamilton (2008) highlight that, whilst there is now generally gender equality between students and instructors, there is still a striking drop in the roles of assistant professors and professors – i.e., a *glass ceiling* in science. The authors analysed more than 100,000 applications between 1996 and 2007 to NIH grant programs to determine whether gender differences occur at some stage of a researcher's career, which may explain the observed attrition. While they found a decrease in female applicants for grants throughout a researcher's career, there is substantial equity of the rates of funded applications between males and females at all stages of the process.

Similar results are observed also by Brouns (2000) on grants awarded to 809 individual applicants by the Dutch Organization for Scientific Research. In this case, however, the stratification by discipline exhibits a higher variability of the success rate for women, from 26% up to 84%, compared to men, from 46% to 76%. Women appear very successful in "hard" sciences (Physics, Mathematics, and Astronomy) and surprisingly unsuccessful in the "soft" natural sciences (Biology, Oceanography, and the Earth Sciences). Bornmann and Daniel (2005) analyse 1954 applications for doctoral and 743 applications for post-doctoral fellowships to a German foundation for the promotion of research in Biomedicine. The odds ratio of the approved doctoral fellowships for females (7%) against males (16%) is found statistically significant after checking for applicant's age, grade, mobility, number of recommendation letters, ratings of reviewers.

Marsh et al. (2008) summarize the major findings of an eight year research program on the analysis of peer reviews in grant applications to the Australian Research Council. Their dataset includes 2331 proposals rated by 6233 external assessors, out of a total of 10023 reviews. They consider issues such as: reliability of reviews, in the sense of an agreement of reviewers across individual proposals and across disciplines; trustworthiness of reviewers nominated by applicants; bias of national reviewers, who give more favorable evaluations than international ones; the positive influence of academic rank, in the sense that professors are more likely to be funded due to their experience and successful research track records; and the positive influence of the prestige of the university affiliation and of the applicant's age. They also consider the influence of an applicant's gender, finding that 15.2% of funded applications were led by females, which was exactly the same percentage as female applicants. Once again, although women are under-represented in the applicants pool, they are equally represented in the funded pool. Their experiments also reject the "matching hypothesis" that reviewers give higher ratings to applicants of their same sex.

Regarding the analytical methodology, research on peer review studies has carried out statistical analyses mainly by means of variants of correlation (Brouns, 2000), Z-tests of proportions (Ley and Hamilton, 2008), regression and more rarely by analysis of variance and discriminant function analysis. Multi-stages peer review processes have been also analysed with latent Markov models (Bornmann et al., 2008). The variants of regression adopted include multiple regression (Wennerås and Wold, 1997; Sandström and Hällsten, 2008), multi-level regression[4] (Jayasinghe et al., 2003; Mutz et al., 2012), logistic regression (Bornmann and Daniel, 2005). The coefficient of the independent variable coding the applicant's gender is taken as a measure of how gender affects the dependent variable, which is typically the score received by the application or its probability (or its logit) of being funded. Other independent variables control factors such as scientific performance, scientific field, age, position, and institution. In this sense, "discrimination is the remaining racial [in our context, gender] difference after statistically accounting for all other race-related [gender-related] influences on the outcome" (Quillian, 2006). However, it is difficult to know that all important characteristics of individuals have been taken into account: a recurring problem known as *the omitted-variable bias*. The inclusion of an omitted control variable may then explain (part of) the remaining gender differences.

2.2. Data mining for discrimination data analysis

Discrimination discovery from data consists in the actual discovery of discriminatory situations and practices hidden in a large amount of historical decision records. The aim is to unveil contexts of possible discrimination on the basis of *legally-grounded* measures of the degree of discrimination suffered by protected-by-law groups in such contexts. The legal principle of under-representation has inspired existing approaches for discrimination discovery based on pattern mining. A common tool for statistical analysis is provided by a 2×2, or 4-fold, contingency table, as shown in Figure 1. Different outcomes between groups are measured in terms of the proportion of people in each group (p_1 for the protected group, and p_2 for the unprotected one) with a specific outcome (benefit denial). Differences and rates of those proportions are commonly adopted as the formal counterpart of the legal principle of group under-representation. They are known in statistics as *risk difference* (RD $= p_1 - p_2$), also known as *absolute risk reduction*; *risk ratio* or *relative risk* (RR $= p_1/p_2$); *relative chance* (RC $= (1 - p_1)/(1 - p_2)$), also known as *selection rate*; *odds ratio* (OR $= p_1(1 - p_2)/(p_2(1 - p_1))$). Starting from a dataset of historical decision records, Pedreschi et al. (2008); Ruggieri et al. (2010a) propose to extract classification rules such as for instance:

```
race=black, purpose=new_car => credit=no
```

[4] In addition to a measurement level random variable, multi-level regression (Goldstein, 2011) includes a subject level random variable modelling variations in a cluster of data. For instance, (Jayasinghe et al., 2003) adopt multi-level regression to take into account correlation in the cluster of ratings of a reviewer, and in the cluster of ratings of a same field of study.

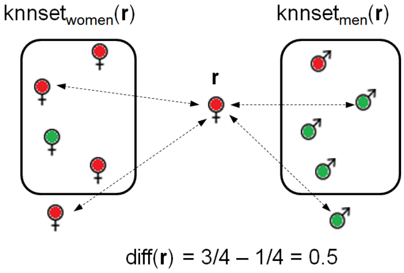
	benefit		
group	denied	granted	
protected	a	b	n_1
unprotected	c	d	n_2
	m_1	m_2	n

$$p_1 = a/n_1 \quad p_2 = c/n_2$$

$$RD = p_1 - p_2 \quad RR = \frac{p_1}{p_2}$$

$$RC = \frac{1-p_1}{1-p_2} \quad OR = \frac{RR}{RC} = \frac{a/b}{c/d}$$

Figure 1: 4-fold contingency table and discrimination measures.


knnset$_{\text{women}}$(**r**)
knnset$_{\text{men}}$(**r**)
r
diff(**r**) = 3/4 – 1/4 = 0.5

```
**Discovery**($t$) {
  for $\mathbf{r} \in \mathcal{P}$ {
    if( $benefit(\mathbf{r}) = denied$
          and $diff(\mathbf{r}) \geq t$ )
      $disc(\mathbf{r}) \leftarrow true$
    else
      $disc(\mathbf{r}) \leftarrow false$
  }
  build a classifier on $\mathcal{P}$
    where the class is $disc$
}
```

Figure 2: Left: example of risk difference $\mathit{diff}(\mathbf{r})$ for $k = 4$. Women are the protected group, $\mathit{knnset}_{women}(\mathbf{r})$ (resp., $\mathit{knnset}_{men}(\mathbf{r})$) is the set of female (resp., male) k-nearest neighbors of $\mathbf{r}$. Red labels benefit denied, green labels benefit granted. Right: pseudo code of k-NN as situation testing. Individuals $\mathbf{r}$ from the protected group $\mathcal{P}$ are first labeled as discriminated or not, and then a classifier is induced for describing those discriminated.

called *potentially discriminatory* rules, to unveil contexts (here, people asking for a loan to buy a new car) where the protected group (here, black people) suffered from under or over-representation with respect to the decision (here, over-representation w.r.t. credit denial). The approach has been implemented by Ruggieri et al. (2010b) on top of an Oracle database by relying on tools for frequent itemset mining. The main limitation of such an approach is that measuring group representation by aggregated values over *undifferentiated* groups results in no control of the characteristics of the protected group, versus or as opposed to others in this context. The high value of a discrimination measure from Figure 1 can be justified by the fact that proportions p_1 and p_2 mix decisions for people that may be very different as per characteristics that are lawful to obtain the required benefit (e.g., skills required for a job position). This results in an overly large number of rules that need to be further screened to filter out explainable discrimination. Luong et al. (2011) overcome this limitation by exploiting the legal methodology of situation testing, which will be presented in Section 2.3.

The approach described so far assumes that the dataset under analysis contains an attribute that denotes the protected group under analysis. The case when data do not contain such an attribute (or it is not even collectable at a micro-data level, e.g., as in the case of sexual orientation) is known as *indirect* discrimination analysis (Ruggieri et al., 2010a), where 'indirect' refers to the exploitation of a known correlation with some other attribute, which can be used as a proxy for group membership. A well-known example is *redlining* discrimination analysis, occurring when the ZIP code of residence is correlated with the race of individuals in highly segregated regions. In this paper, we restrict to consider direct discrimination analysis.

Finally, we mention the related research area of *discrimination prevention* in data mining and machine learning (Calders and Verwer, 2010; Kamiran and Calders, 2012; Hajian and Domingo-Ferrer, 2012), where the problem is to design classification algorithms that trade off accuracy for non-discrimination in making predictions. Discriminatory predictions may be the result of a bias of the classifier induction algorithm, or of learning from training data traditional prejudices that are endemic in reality. Summaries of contributions in discrimination discovery and prevention are collected in a recent book (Custers et al., 2013). In particular, Romei and Ruggieri (2013) present, in a chapter of that book, a multi-disciplinary annotated bibliography of statistical tools, legal argumentations, economic models, and experimental approaches for discrimination data analysis.

2.3. Situation testing and k-NN

In a legal setting, situation testing is a quasi-experimental approach to investigate for the presence of discrimination by checking the factors that may influence decision outcomes. Pairs of research assistants, called *testers*, undergo the same kind of selection. For example, they apply for the same job, they present themselves at the same night club, and so on. Within each pair, applicant characteristics likely to be related to the situation (characteristics related to a worker's productivity on the job in the first case, look, age and the like in the second case) are made equal by selecting, training, and credentialing testers to appear equally qualified for the activity. Simultaneously, membership to a protected group is experimentally manipulated by pairing testers who differ in membership – for example, a black and a white, a male and a female, and so on. Observing significant difference in the selection outcome between testers is a *prima facie* evidence of discrimination, i.e., a proof that, unless rebutted, would be legally sufficient to prove the claim of discrimination. For applications of situation testing, we refer to Bendick (2007), covering employment discrimination in the US; to Rorive (2009), covering the EU member States context; and to Pager (2007), including an appendix on the design of situation testing experiments.

In Luong et al. (2011), the idea of situation testing is exploited for discrimination discovery just inverting the

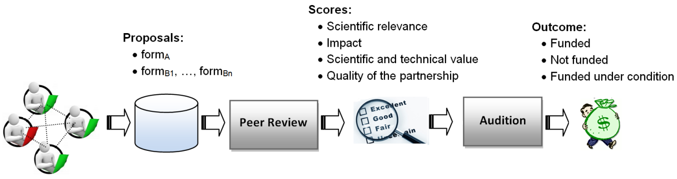


Figure 3: The two-phases review process of the FIRB "Future in Research" call.

point of view. Given past records of decisions taken in some context, for each member of the protected group with vector of attributes $\mathbf{r}$ suffering from a negative decision outcome (someone who may claim to be a victim of discrimination), we look for $2k$ testers with similar characteristics. Such characteristics are legally admissible in affecting the decision, apart the one of being or not in the protected group. Similarity is modeled via a distance function between tuples. If we can observe significantly different decision outcomes between the k-nearest neighbors of $\mathbf{r}$ belonging to the protected group and the k-nearest neighbors belonging to the unprotected group, we can ascribe the negative decision to a bias against the protected group, hence labeling the individual $\mathbf{r}$ as discriminated. This approach resembles the k-nearest neighbor (k-NN) classification model, where the class of an individual is predicted as the most frequent class among its k-nearest neighbors. Difference in decision outcomes between the two groups of neighbors is measured by any of the functions from Figure 1, calculated over the proportions for the two sets of testers. Throughout the paper, we consider risk difference $\mathit{diff}(\mathbf{r}) = p_1 - p_2$ with the intuitive reading that it represents the difference in the frequency p_1 of negative decisions in the neighbors of the protected group with respect to frequency p_2 in the neighbors of the unprotected group (see Figure 2 for an example). A value $\mathit{diff}(\mathbf{r}) > t$, for a maximum threshold t, implies that the negative decision for $\mathbf{r}$ is not explainable on the basis of the (legally-grounded) attributes used for distance measurement, but rather it is biased by group membership. $\mathit{diff}(\mathbf{r})$ is then a measure of the strength of such a bias. When $\mathit{diff}(\mathbf{r}) > t$, individual $\mathbf{r}$ is labeled as discriminated by setting a new attribute *disc* to *true*. Starting from this labeling procedure, the actual learning of the conditions of discrimination is then modeled as a standard classification problem, where the class is the attribute *disc*. The overall procedure is reported in Figure 2.

3. Case study: data understanding & preparation

In this section, we start the analysis of the case study of an Italian national call for research projects. We introduce the call and its evaluation process, the available data, and the features selected to form the dataset in input to the discrimination discovery analysis.

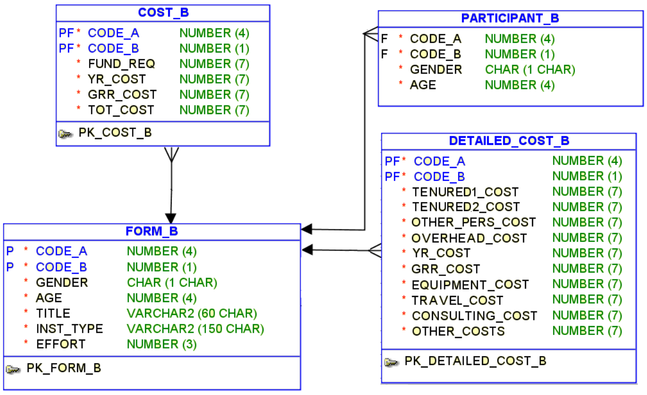


Figure 4: Input data on research units.

3.1. The FIRB "Future in Research" call

In 2008, the Italian Ministry of University and Research published a call for scientific research projects under the Basic Research Investment Fund (FIRB) reserved to young scientists – the *FIRB "Future in Research"* call. The scientific scope of the call is very broad, ranging from social sciences and humanities, to physical sciences and engineering, to life sciences. Research proposals are submitted by a consortium of one or more research units, with a *principal investigator* (from now on, PI) and zero or more *associate investigators* heading each unit and affiliated to an Italian university or to a public research organization. Research proposals are distinguished in two programs, depending on whether the PI holds a non-tenured position and she/he is at most 33 years old at the time of the call (program P_1), or she/he holds a tenured position and she/he is at most 39 years old (program P_2). Each program has its own total budget, but the submission forms and the evaluation procedures are the same for both.

The submitted proposals consist of a description of work and a budget for each research unit, called the *B forms*, and of a description of work and a budget for the whole proposal, called the *A form*. The global budget is basically the sum of the budgets of the research units participating in the project. The A form also contains the curriculum vitae of the PI, a list of her/his main publications, and an abstract of the proposal. The hiring of at least one *young researcher* (defined as "a post-doc or a post-degree of at most 32") per project proposal is required by the call. Invitation of good reputation researchers from abroad to spend some period working on the project is instead an option.

Project proposals underwent a two-steps evaluation process, as shown in Figure 3. The first step consisted in a blind peer-review by national and international reviewers resulting in four scores about:

(S1) scientific relevance of the proposal (score 0–8);

(S2) impact of the proposal (score 0–7);

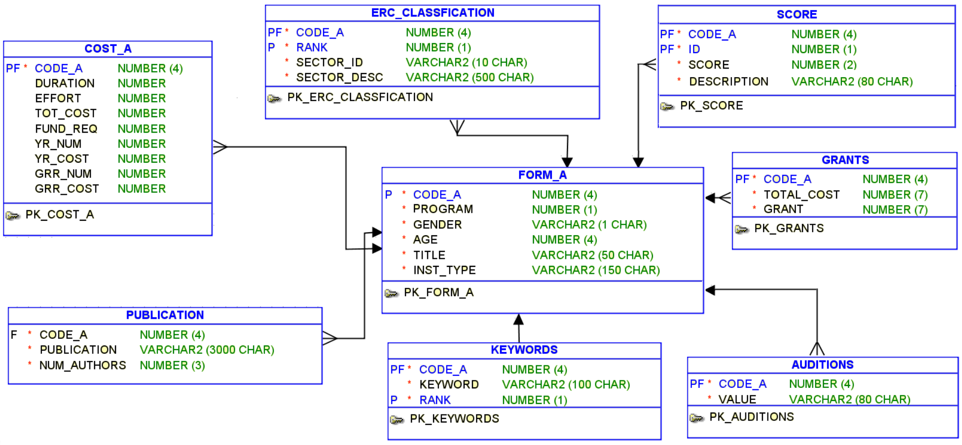


Figure 5: Input data on research proposals and evaluation results.

(S3) scientific and technical value of the proposal (score 0–15);

(S4) quality of the partnership (score 0–10).

Only project proposals that received the best score for all of the four evaluation criteria (i.e., a total score of 40) were admitted to the second step, which consisted in an audition of the PI in front of a panel of national experts. The panel ranked the proposals into three classes: "to be funded", "to be funded if additional budget were available"[5] and "not to be funded". In the first two cases, the panel also decided a budget cut with respect to the budget requested.

3.2. Data sources

Anonymized data on project proposals and evaluation results were made available to us as an Oracle relational database. Proposals are identified by unique IDs `CODE_A`. Similarly, research units have unique IDs `CODE_B`.

Data on research units (see Figure 4) are retrieved from B forms of project proposals. Table `FORM_B` contains, for each research unit, attributes on the associate investigator leading the unit (gender, age, title, institution) and on the planned effort of the research unit in person/months. For each research unit, three other groups of data are available:

- participants to the research unit, whose gender and age attributes are stored in the `PARTICIPANT_B` table;
- detailed costs of the research unit, stored in the `DETAILED_COST_B` table, including costs of: tenured personnel, personnel to be hired, equipment, overhead, travel and subsistence, consulting and other costs;
- aggregated costs, stored in the `COST_B` table, including: the costs of hiring young researchers and good reputation researchers, the total budget of the research unit, and the eligible costs to be funded by the call.

Data on research proposals (see Figure 5) are retrieved from A forms. Table `FORM_A` contains data on the PI (gender, age, title, institution) and the research program of the proposal (P_1 or P_2). A few auxiliary tables follow:

- `PUBLICATION` stores the list of publications of the PI. Authors' names have been removed, but the number of authors is recorded;
- `ERC_CLASSIFICATION` stores the scientific area of the research project according to the European Research Council (ERC) classification; more than one area could have been chosen for a research proposal, e.g., in case of multi-disciplinary topics, with the first one representing the main area[6];
- `KEYWORDS` and `ABSTRACT` store respectively an ordered list of keywords and the textual abstract of the proposal, in English;
- `COST_A` stores the duration of the project in months, and aggregated budget data: total effort in person/-months, total cost of the project, total eligible costs to be funded, number of young researchers and their total cost, number of good reputation researchers and their total cost.

Data on the evaluation results is shown in the rightmost tables of Figure 5. The scores obtained by a proposal over the four evaluation criteria of the first step of the evaluation process are stored in the `SCORE` table. Each criterium is coded with an ID. The ranks assigned by the commission of national experts to proposals in the second step of the evaluation process is stored in the `AUDITIONS` table. For proposals ranked as "to be funded" or "to be funded if additional budget were available", the total cost and grant assigned to the project after budget cut is stored in the `GRANTS` table.

[5] At that time, an increase of the budget of the call was under consideration.

[6] The main area is used in the first step of the evaluation process to select the peer-reviewers of the proposal from a pool of area experts.

Name	Description	Type
Features on principal and associate investigators		
gender	Gender of the PI	Nominal
region	Region of the institution	Nominal
city	City of the institution	Nominal
inst_type	Type of the institution	Nominal
title	Title of the PI	Nominal
age	Age of the PI	Numeric
pub_num	No. of publications of the PI	Numeric
avg_aut	Average number of authors in pubs	Numeric
f_partner_num	No. of females among principal or associate invest.	Numeric
Project costs (absolute values in €)		
tot_cost	Total cost of the project	Numeric
fund_req	Requested grant amount	Numeric
fund_req_perc	fund_req over tot_cost	Numeric
yr_num	No. of young researchers	Numeric
yr_cost	Cost of young researchers	Numeric
yr_perc	yr_cost over tot_cost	Numeric
grr_num	No. of good reputation researchers	Numeric
grr_cost	Cost of good reputation researchers	Numeric
grr_perc	grr_cost over tot_cost	Numeric
Research areas		
program	Program P_1 or P_2 of the proposal	Nominal
d1_lv1, d2_lv1, d3_lv1	1^{st}, 2^{nd} and 3^{rd} area at the 1^{st} ERC level	Nominal
d1_lv2, d2_lv2, d3_lv2	1^{st}, 2^{nd} and 3^{rd} area at the 2^{nd} ERC level	Nominal
d1_lv3, d2_lv3, d3_lv3	1^{st}, 2^{nd} and 3^{rd} area at the 3^{rd} ERC level	Nominal
Evaluation results		
S1	Score S1 assigned by the peer-reviewer	Numeric
S2	Score S2 assigned by the peer-reviewer	Numeric
S3	Score S3 assigned by the peer-reviewer	Numeric
S4	Score S4 assigned by the peer-reviewer	Numeric
peer-review	Passed or rejected at the peer-review	Nominal
audition	Passed or rejected at the audition (i.e., proposal funded)	Nominal
grant	Amount granted after budget cut	Numeric

Table 1: Attributes of the dataset of the case study.

3.3. Data preparation

The data preparation phase produced a dataset for the discrimination analysis in the form of a single relational table, including both source and derived features for each project proposal. Table 1 summarizes four groups of features.

Features on the principal and associate investigators. These include gender, age, and title of the PI; number of publications and average number of authors in publications of the PI; region (North, Center, South of Italy), city and type of her/his institution (University, Consortium or Other); and number of female principal or associate investigators in the project proposal.

Project costs. Several costs are considered: total cost of the project, requested grant (both absolute and in proportion to the total cost), number and cost of young researchers, number and cost of good reputation researchers.

Research areas. In addition to the research program a proposal is submitted to (P_1 or P_2), up to three research areas are included, the first of which is the main area, according to the ERC classification. Such a classification consists of a three-level hierarchy. The top level includes Social sciences and Humanities (SH), Physical sciences and Engineering (PE), and Life Sciences (LS). The second and third levels (coded, e.g., as PE_*n* and PE_*n*_*m*) include 25 and 3792 sub-categories respectively.

Evaluation results. The following attributes are included: the scores (S1)-(S4) received at the peer-review, whether the project passed the first evaluation phase (i.e., the peer-review), whether the project passed the second evaluation phase[7] (i.e., the audition), the actual amount granted after budget cut.

4. Case study: risk difference analysis

Since research proposals of programs P_1 and P_2 are evaluated in isolation (due to distinct budgets), from now on, we act as if there were two datasets, one per program. Program P_1 received 1804 applications, 923 of which are from female PIs; program P_2 received 1986 ones, 792 of which from female PIs.

4.1. Exploratory data analysis of gender differences

Table 2 summarizes the proportion of genders in the two phases of the evaluation process: peer-review and audition. It is readily checked that, for both programs, the proportion of female PIs progressively decreases when moving from applicant proposals to proposals passing the peer-review up to those passing the audition decision. Let us quantify such a decrease by means of discrimination measures. Figure 6 shows the 4-fold contingency tables of

[7] Since no additional budget was available for the call, proposals ranked as “to be funded if additional budget were available” are considered as not passing the audition.

	PIs		Peer-review passed		Audition passed	
	Male	Female	Male	Female	Male	Female
P_1	881 (48.8%)	923 (51.2%)	43 (58.1%)	31 (41.9%)	25 (59.5%)	17 (40.5%)
P_2	1194 (60.1%)	792 (39.9%)	100 (76.3%)	31 (23.7%)	51 (81%)	12 (19%)

Table 2: Aggregate data on gender differences.

	Peer-review P_1		
Applic.	Rejected	Passed	
Female	892	31	923
Male	838	43	881
	1730	74	1804

$p_1 = 892/923 = 0.966$
$p_2 = 838/881 = 0.951$
$RD = 0.015 \quad RR = 1.02$
$RC = 0.69 \quad OR = 1.48$

Audition P_1		
Rejected	Passed	
14	17	31
18	25	43
32	42	74

$p_1 = 14/31 = 0.452$
$p_2 = 18/43 = 0.419$
$RD = 0.033 \quad RR = 1.08$
$RC = 0.94 \quad OR = 1.15$

	Peer-review P_2		
Applic.	Rejected	passed	
Female	761	31	792
Male	1094	100	1194
	1855	131	1986

$p_1 = 761/792 = 0.961$
$p_2 = 1094/1194 = 0.916$
$RD = 0.045 \quad RR = 1.05$
$RC = 0.46 \quad OR = 2.24$

Audition P_2		
Rejected	passed	
19	12	31
49	51	100
68	63	131

$p_1 = 19/31 = 0.613$
$p_2 = 49/100 = 0.49$
$RD = 0.123 \quad RR = 1.25$
$RC = 0.76 \quad OR = 1.65$

Figure 6: 4-fold contingency tables and discrimination measures.

passing the peer-review and of being funded for proposals in programs P_1 and P_2. Consider first the peer-review phase. Recall that the measures of risk difference (RD) and risk ratio (RR) compare the proportions of *rejected* proposals. Due to the small fraction of projects passing the phase, it turns out that RD and RR cannot highlight differences in the outcomes. Overall, the vast majority of proposals were rejected. In fact, RR is only 1.02 for P_1 and 1.05 for P_2; RD is only 1.5% for P_1, and a modest 4.5% for P_2. On the other hand, since relative chance (RC) compares the *success* rates, it highlights major differences: the chance of passing the peer-review for a female is only 69% of the chance of a male for program P_1, and only 46% for program P_2. Finally, since the odds ratio (OR) is the ratio between RR and RC, it highlights differences in both rejection and success rates. Consider now the audition phase. Rejected and funded projects are now more evenly distributed. The discrimination measures highlight no significant difference for program P_1. Differences in program P_2 are lower than the first phase, yet still moderately high.

Are the different success rates of males and females due to legitimate characteristics or skill differences of the gender of applicants? In order to answer such a question, Figure 7 reports the distributions of the age of the PIs, of the number of her/his publications, and of a few costs of project proposals (young researchers, good reputation researchers, total cost, request grant). Distributions are distinguished for gender of the PI and for program of the project proposal. The distribution of age across gender highlights no difference for both programs. Notice that the distributions across programs are clearly distinguished due to the requirements of each program in the call for proposals. However, the plot of the number of publications shows that males have a slightly higher productivity than females, for both programs P_1 and P_2. As an example, about 37% of females in program P_2 have more than 20 publications, against a percentage of 48% for males. Turning our attention on project costs, we observe that proposals led by females require slightly lower costs for young researchers than proposals led by males, in both programs. The situation is similar for the total cost and the requested grant: the average total cost is € 980K for females and € 1080K for males. The distributions of costs for good reputation researchers are, instead, similar. Notice that such costs are non-zero for only 19% of proposals in program P_1 and only 10% in P_2.

Summarizing, even though an analysis of distributions provides some hints on gender differences, it is still too gross grained to draw any conclusion about the presence of discrimination. Aggregations at the level of the whole dataset may hide differences in smaller niches of data. Unveiling these niches is precisely the objective of the discrimination discovery analysis.

4.2. Distributions of gender risk difference

Let us instantiate the approach of situation testing (see Section 2.3) by exploring risk differences. Let $\mathbf{r}$ be a project proposal led by a female PI that did not pass the peer-review phase. The function $\mathit{diff}(\mathbf{r}) = p_1 - p_2$ measures the risk difference between the rejection percentage p_1 of its k-nearest neighbor proposals headed by female PIs and the rejection percentage p_2 of its k-nearest neighboring proposals headed by male ones. Distance is measured on the basis of proposal's characteristics that are (legally) admissible in affecting the (first or second phase) decision. We consider here all the features of Table 1 apart from the project evaluation results and the gender of the PI. Similarity is modeled via the distance function adopted in the experiments by Luong et al. (2011), which consists of the Manhattan distance of z-scores for continuous attributes in $\mathbf{r}$, and of the percentage of mismatching attributes for discrete ones. The higher $\mathit{diff}(\mathbf{r})$ is, the more the negative decision on proposal $\mathbf{r}$ is unexplainable by differences in the compared characteristics. The residual explanation is then the gender of the PI, which implies a *prima facie* evidence of gender discrimination, or the lack of further explanatory variables – the *omitted variables.* A critical choice concerns how to set the k constant. Figure 8 (a,b) shows the distributions of $\mathit{diff}()$ for $k = 4, 8, 16, 32$ with reference to proposals from programs P_1 and P_2. As k increases, the distributions tend to flatten (for k sufficiently large, the risk differences of all proposals collapse to a unique value). From now on, we fix $k = 8$, which means comparing each proposal with 0.9% of the proposals in program P_1 (= 16/1804, where 16 is $2k$, and 1804 is the overall number of proposals), and with 0.8% of the proposals in program P_2.

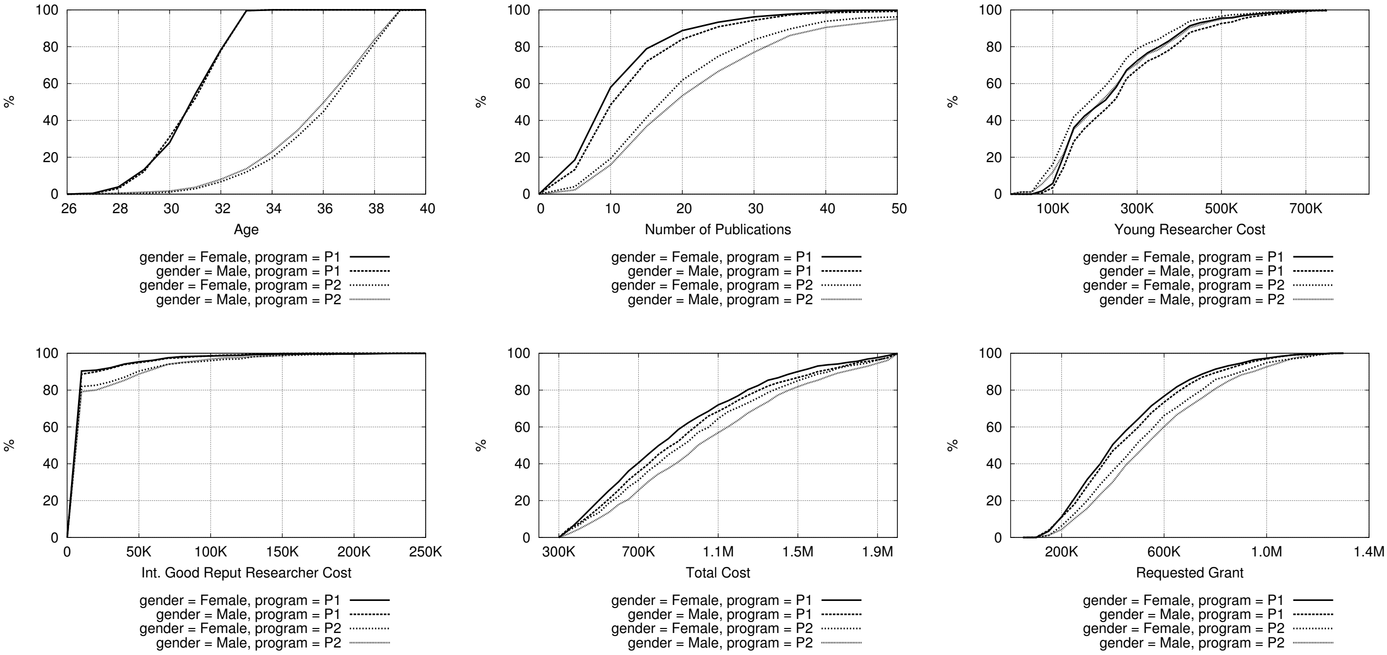
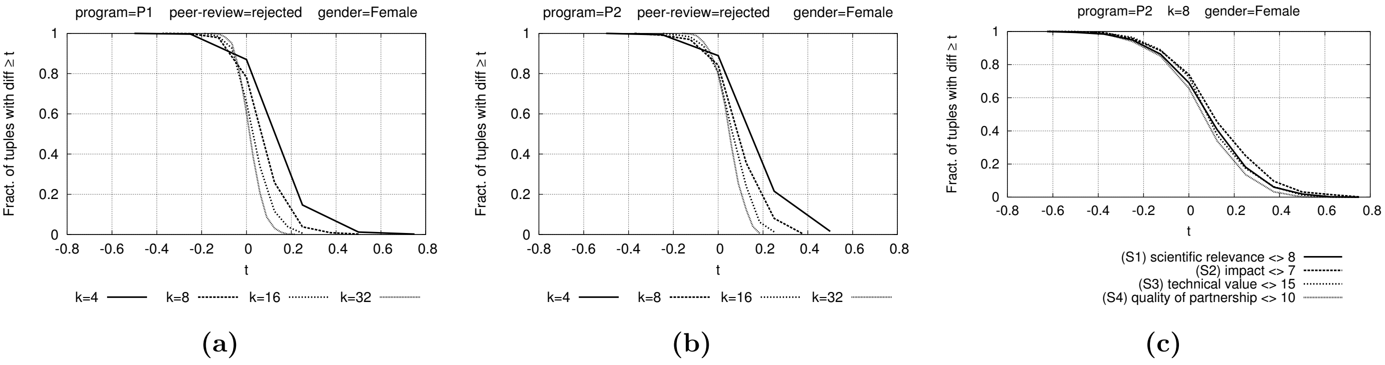


Figure 7: Cumulative distributions across gender of PIs and program of proposals.



(a) **(b)** **(c)**

Figure 8: Cumulative distributions of *diff*().

Recall that only the project proposals that receive the highest scores (S1)-(S4) pass the peer-review phase. It is interesting then to look at the distributions of *diff*() separately for each score. This is shown in Figure 8 (c), where the "benefit denied" decision is set as receiving a score lower than the maximum. The impact of the project (S2) appears as the most biased criteria.

Distributions might also unveil forms of multiple discrimination. Figure 9 shows the distributions of risk difference for two possibly discriminated groups in isolation, namely female PIs and PIs affiliated to institutions from the South of Italy (an historically disadvantaged region of Italy), and for PIs belonging to both groups. The two groups in isolation exhibit some risk difference, with gender bias being more prominent than bias against people from the South. However, no increase in risk difference can be observed for the sub-group of female PIs from the South when compared to the whole group of female PIs.

5. Case study: discrimination model extraction

In this section, we start applying a discrimination discovery approach to the pre-processed datasets of proposals in program P_1 (resp., P_2) with reference to the peer-review decision. We will not be considering the decision of the audition phase due to three motivations. *First*, the number of proposals involved in the second phase of the reviewing process is rather small, hence we run the risk of drawing no (statistically) significant conclusion. *Second*, the discrimination measures in Figure 6 highlight higher gender differences in the peer-review decisions than in the audition decisions, so we expect higher chances of finding non-negligible contexts of clear discrimination. *Third*, and most important, the set of features available in Table 1 appear adequate as control factors for the decision of the first phase only. In fact, peer-reviewers had access to the proposal text, to the curriculum and list of publications

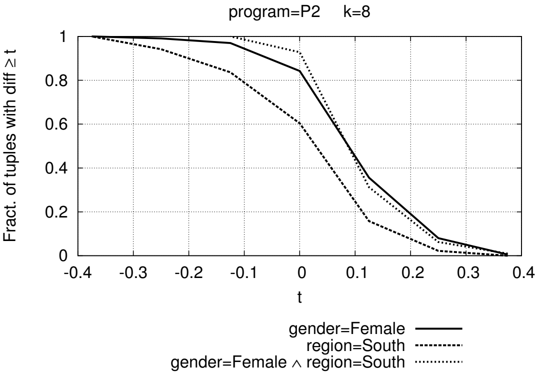


Figure 9: Cumulative distributions of *diff* ().

of the PI, and to the budget data. This is about the set of features listed in Table 1. On the contrary, the panel of national experts "entered in personal contact" with the PIs during the auditions, so their decisions are affected by additional factors not recorded in the data, e.g., physical characteristics of the PI, proficiency in speaking, motivation, and appropriateness of answers to questions. The omitted-variable bias in analysing data with reference to the decision of the panel of national experts would then be considerably high. Consequently, any finding of possible discrimination would be questionable.

5.1. Before data mining: what can regression tell us?

Data analysts from economic and social sciences have typically adopted logistic regression as a tool for testing an hypothesis of possible discrimination. Before starting our data mining analysis, let us then follow such an approach and discuss what conclusions can be made *without* applying data mining methods. Logistic regression is a form of multiple linear regression:

$$logit(P(Y=1)) = \alpha + \sum_{i=1}^{N} \beta_i X_i$$

where the logit of the dependent variable value $Y = 1$ is estimated as a linear function of the independent variables $X_1, \dots, X_n$. The logit function $logit(P(Y = 1)) = log(P(Y = 1)/(1 - P(Y = 1)))$ is the log of the odds of the probability $P(Y = 1)$. By exponentiating the equation sides, we obtain:

$$\frac{P(Y=1)}{1-P(Y=1)} = e^{\alpha+\sum_{i=1}^{N}\beta_i X_i} = e^{\alpha}\prod_{i=1}^{N} e^{\beta_i X_i}$$

The value β_i can then be interpreted as the variation coefficient of the logarithm of the odds of the event $Y = 1$ due to a linear variation of the factor X_i, all other control factors being constant. A nominal feature X with values $v_1, \dots, v_k$ is modeled in this framework by $k - 1$ independent indicator variables $X = v_1, \dots X = v_{k-1}$. The coefficients of these features model the variation of the logit of $P(Y = 1)$ with respect to the default value $X = v_k$.

	Model for P_1		Model for P_2		
Variable	Coeff.	(Std. err)	Coeff.	(Std. err)	
gender = Female	-0.33	(-0.39)	-0.87	(0.30)	[***]
region = North	0.14	(0.30)	0.22	(0.22)	
region = South	-0.02	(0.35)	-0.42	(0.28)	
inst_type = Univ	-0.42	(0.62)	-0.43	(0.40)	
inst_type = Cons	-0.36	(0.50)	0.04	(0.45)	
age	-0.03	(0.09)	0.01	(0.04)	
pub_num	0.01	(0.01)	-0.01	(0.01)	
avg_aut	0.01	(0.01)	0.03	(0.02)	
f_partner_num	-0.02	(0.25)	0.12	(0.16)	
tot_cost	0.10	(0.35)	0.03	(0.26)	
fund_req	-0.13	(0.50)	-0.05	(0.37)	
fund_req_perc	0.51	(0.37)	0.39	(0.28)	
yr_num	-0.11	(0.26)	-0.56	(0.19)	[***]
yr_cost	-0.09	(0.35)	-0.03	(0.26)	
yr_perc	0.31	(0.26)	0.27	(0.19)	
grr_num	-0.28	(0.46)	0.22	(0.24)	
grr_cost	-0.09	(0.35)	-0.03	(0.26)	
grr_perc	0.16	(0.32)	0.26	(0.21)	
d1_l1 = PE	-0.42	(0.33)	-0.34	(0.25)	
d1_l1 = SH	0.08	(0.33)	0.03	(0.29)	

Table 3: Logistic regression models for datasets of proposals for programs P_1 and P_2. The dependent variable is *peer-review = passed*. Coefficients marked by [***] are statistically significant at the 99% confidence level.

Table 3 shows logistic regression models for the datasets of proposals in program P_1 and P_2. The event $Y = 1$ is here *peer-review = passed*. Standard errors and statistical significance of regression coefficients are also shown. In both models, the regression coefficient of the indicator variable *gender = Female* is negative, which means that, all other factors being equal, female PIs have lower odds of passing the peer-review phase: by a factor of $e^{-0.33} = 0.72$ for program P_1, and of $e^{-0.87} = 0.42$ for program P_2 w.r.t. the odds of male PIs. For program P_2, the null hypothesis that the coefficient is zero is rejected at the 99% level of statistical significance. The region of the institution of the PIs affects the odds of passing the peer review as well, particularly for proposal in program P_2: PIs from the North of Italy have higher chances, whilst those from the South have lower ones. Variables on age, number of publications, average number of authors in publications have coefficients close to zero. Concerning cost variables, proposals with higher percentage of costs covering young researchers (variables *yr_perc*, *fund_req_perc*) have higher odds of passing the peer-review. This is not unexpected, since the call explicitly states the objective of funding start-up research groups of young researchers. However, proposals with large (yet, young) groups (variable *yr_num*) are disfavored in the peer-review decision. Moreover, competition appears to be harder in the area of Physical sciences and Engineering (PE) rather than in Humanities (SH), and Life Sciences (LS); and for PIs from the University (variable *inst_type = Univ*) compared to PIs from other institutions. Finally, the literature on discrimination analysis accounts for an *included-variable bias* (Killingsworth, 1993), namely for control variables that incorporate some form of gender discrimination. One such variable is *f_partner_num*, i.e., the number of female principal or associate investigators. Since its coefficients

				Accuracy		Precision		Recall		f-Measure		Size	
Id	Set	Alg	CS	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2	P_1	P_2
1	D_{13}	Jrip	Y	48.4	45.2	33.1	40.2	**97.0**	93.4	0.49	0.56	35	23
2	D_{19}	C4.5	N	77.6	70.7	54.3	58.7	83.1	74.5	0.66	0.66	302	184
3	D_{19}	C4.5	Y	56.4	53.0	36.6	44.0	93.5	92.3	0.53	0.60	78	73
4	D_{19}	Jrip	Y	53.5	45.4	35.4	40.4	**97.0**	95.5	0.52	0.57	21	11
5	D_{19}	Part	N	72.3	66.6	47.9	54.5	77.9	67.8	0.59	0.60	74	33
6	D_{19}	Part	Y	61.9	57.3	39.7	46.4	90.9	88.8	0.55	0.61	74	87
7	D_{27}	C4.5	N	**82.2**	**74.5**	**62.3**	**63.4**	78.8	75.9	**0.7**	**0.69**	2051	4645
8	D_{27}	C4.5	Y	50.7	37.6	33.5	37.6	92.2	100	0.49	0.55	9	325
9	D_{27}	Jrip	Y	50.0	46.5	33.7	41.0	96.1	**96.9**	0.5	0.58	12	11
10	D_{27}	Part	Y	61.2	53.1	38.1	43.7	80.1	86.7	0.52	0.58	128	113

Table 4: Top 10 classification models of discriminated proposals.

are small, it appears that gender discrimination bias, if present, is directed mainly against the PI, and not against the group of investigators.

The conclusions drawn from Table 3 are certainly more informative than the explorative data analysis of Section 4. However, whilst they reveal some gender bias at the level of the whole datasets, there is no indication of whether this is uniformly distributed or whether there are some contexts with a very high bias. Unfortunately, the statistical regression approach is limited to the verification of an hypothesis. Thus, one should explicitly figure out a possible context and re-compute a regression model for proposals in such a context. The purpose of our data mining approach is precisely to let such contexts *emerge* as a result of the analysis.

5.2. A classification model of the discriminated proposals

The number of proposals led by female PIs that did not pass the peer-review phase amounts at 892 for program P_1 and at 761 for program P_2. Now, we intend to extract from these two datasets a global description of proposals whose negative decision is discriminatory according to the legal methodology of situation testing. We proceed as follows. First, we set a threshold value t to the maximum admissible risk difference. Values of risk differences greater than 0.05 (i.e., 5%) have been considered *prima facie* evidence of discrimination in some legislations and law cases, and Figure 8 (a) supports this choice in practice. In order to be more stringent, we assume from now on the higher threshold $t = 0.10$. Second, a proposal $\mathbf{r}$ is labeled as discriminated if its risk difference is greater or equal than t – technically, we introduce a binary attribute *disc* defined as $disc(\mathbf{r}) = true$ iff $diff(\mathbf{r}) \geq 0.10$. These two steps allow us for reducing the problem of characterizing discriminatory decisions to the standard problem of inducing a classification model, where the class attribute is the newly introduced attribute *disc*. The resulting datasets have a distribution of *disc* = *true* and *disc* = *false* values of 26-74% for program P_1, and of 38-62% for P_2.

Since the intended use of the classification models is to *describe* global conditions under which a proposal led by a female PI is rejected at the peer-review phase with a risk difference of 0.10 or above, we restrict the search space to classification models that are readily interpretable (e.g., before a court in a law case). We experimented classification rule models (RIPPER by Cohen (1995), and PART by Frank and Witten (1998)) and decision trees (C4.5 by Quinlan (1993)). Classification models are evaluated by the objective interestingness measures of accuracy, precision, recall and f-measure for the class value *disc* = *true* using a 10-fold cross validation. The actual classification model is extracted from the whole dataset. Other settings that have been experimented are mainly concerned with tackling the unbalanced distribution of class values, and they include standard approaches: uniform resampling of the training folds, cost-sensitive induction of classifiers,[8] and meta-classification approaches (bagging and boosting). We relied on the Weka tool by Witten and Frank (2011) for algorithms and as experimental environment. Finally, we also varied the set of predictive attributes in order to evaluate the explanatory power of different subsets:

- set D_{13} includes a subset of features of the PI (*age*, *title*, *pub_num*, *avg_aut*), of project costs (*yr_num*, *yr_cost*, *grr_num*, *grr_cost*, *tot_cost*, *fund_req*), of the research area (*d1_lv1*, *d1_lv2*) as well as the class attribute *disc*;
- set D_{19} adds features of project costs (*fund_req_perc*, *yr_perc*, *grr_perc*), of the PI (*inst_type*, *region*) and of the participants (*f_partern_num*);
- set D_{27} also includes the remaining attributes of the ERC hierarchy and the attribute *city* of the PI.

Table 4 reports the top 10 classification models obtained. For each model, the table includes: the set of predictive attributes, the extraction algorithm, whether cost-sensitive (CS) classification is adopted, and performance measures for both program P_1 and P_2. All of the top 10 classifiers use resampling, whilst none of them adopt meta-classification. The size of a model measures its structural complexity: for classification rules, it is the number of rules; for decision trees, it is the number of leaves.

A few comments follow on the lessons learned in tuning models and parameters. *First*, resampling of the training

[8] The best performance is obtained with a cost of misclassifying *disc* = *true* set to 2.5 times the cost of misclassifying *disc* = *false*.

folds reveals itself as an effective technique, improving performances both in term of accuracy and f-measure, and irrespectively of the model type and subset of attributes. Using in addition cost-sensitive classification does not improve further. Compare for instance rows 2 vs 3, 5 vs 6, 7 vs 8 from Table 4, where the only difference between the pairs is in the use or not of misclassification costs. *Second*, the impact of the set of predictive attributes is dependent on the classification model. Jrip and PART both benefit from larger sets as per accuracy and f-measure when moving from D_{13} to D_{19}, but then the additional attributes in D_{27} worsen the performances. Contrast for example rows 1 vs 4 vs 9, and 6 vs 10. This also holds for C4.5 models when using misclassification costs (see rows 3 vs 8). However, when using resampling only, there is an improvement from D_{19} to D_{27} (rows 2 and 7). C4.5 with resampling on D_{27} (row 7) turns out to be the best model with respect to both accuracy and f-measure. *Third*, we highlight the importance of extracting models that trade off performance with simplicity. The best model (row 7) is, unfortunately, the most complex one. The *global* description of discriminatory decisions it provides is accurate but sparse in too many conditions, whose validation, e.g., by a legal expert, is impractical. This motivates the search for a few *local* contexts of possible discrimination.

6. Case study: rule reasoning and validation

In this section, we report four rules filtered out from the top 10 classifiers. They have been ranked in the top positions on the basis of both objective measures (precision, recall, average *diff*(), odds ratio) and subjective ones (interpretability, relation with known stereotypes). The antecedent of a rule unveils a context of *prima facie* evidence of gender discrimination. Proposals led by female PIs in such a context observe different decisions (with risk difference of at least 0.10) of peer-reviewers between projects with similar characteristics led by male PIs and projects with similar characteristics led by female PIs. We validate the statistical significance of such a context by means of logistic regression. This way, we merge the capability of the k-NN as a situation testing approach for *discovering* contexts of possible discrimination with the capability of statistical regression for *testing* hypothesis of possible discrimination – thus obtaining the best of the two worlds.

A technical note is in order. For each of the four rules, all proposals led by female PIs in the context of the rule result to have been rejected at the peer-review. As a consequence, the coefficient of the independent variable *gender = Female* in a logistic regression model cannot be calculated – this is known as the *separation problem* (Heinze, 2006). We will then apply the *Firth* logistic regression (Firth, 1993), also called penalized maximum likelihood method, which takes into account such a problem. For the same reason, we will calculate the odds ratio (OR) of a rule by applying the plus-4 correction, which consists of adding a fictitious +1 to each cell in the contingency table of Figure 1.

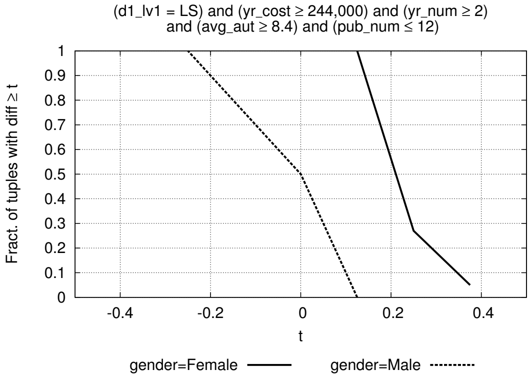


Figure 10: Cumulative distributions of *diff*() for proposals satisfying the antecedent of rule R1.

6.1. Rule R1: life sciences in program P_1

The first rule concerns proposals in program P_1. It highlights a possible discrimination against female PIs in the area of Life Sciences (LS).

```
R1: (d1_lv1 = LS) and (yr_num >= 2) and
    (yr_cost >= 244,000) and (pub_num <= 12) and
    (avg_aut >= 8.4) => disc=yes
    [prec=1.0] [rec=0.095] [diff=0.165] [OR=11.14]
```

The antecedent of the rule points out a context of research proposals requiring two or more young researchers, having a cost for them of € 244K or more, and such that the PI has at most 12 publications with a mean number of authors of 8.4 or more. There are 33 proposals in the contex: 8 led by male PIs, 2 of which passed the peer-review, and 25 led by female PIs, none of which passed the peer-review. All of the 25 proposals led by female PIs have been labeled as discriminated, i.e., the precision of the rule is 100%. Among the proposals led by female PIs that are labeled as discriminated, 9.5% satisfy the antecedent of the rule, i.e., the recall of the rule is 9.5%, which makes the context rather relevant for the anti-discrimination analyst. With reference to proposals of the LS area only, recall lifts up to 27%. The average risk difference measure of the 25 proposals led by female PIs is 16.5%, which is much higher than the discrimination threshold value of 10%. Figure 10 reports the cumulative distributions of *diff*() for proposals satisfying the antecedent of the rule R1 distinguishing female and male led projects. This is more informative than simply the average risk difference. Moreover, it highlights the other side of discrimination, namely *favoritism*: proposals led by males exhibit very low or even negative risk differences, or stated otherwise, they have been favored in comparison to similar projects led by female PIs. Finally, the (corrected) odds risk is 11.14: the odds of proposals led by female PIs of being rejected at the peer-review is 11.14 times the odds of those led by male PIs.

Rule R1 unveils then a possible gender discrimination in the Life Science area when proposals are ambitious (more

Variable	Rule R_1		Rule R_2		Rule R_3		Rule R_4	
	Coeff.	(Std. err)	Coeff.	(Std. err)	Coeff.	(Std. err)	Coeff.	(Std. err)
gender = Female	-1.64	(1.40) [*]	-1.37	(1.38) [**]	-0.86	(1.22)	-1.20	(0.63) [***]
inst_type = Univ	0.11	(2.41)					-1.58	(1.28)
inst_type = Cons.							-0.70	(1.63) [**]
age	-0.05	(0.46)	0.03	(0.42)	0.01	(0.44)	0.01	(0.20)
pub_num	0.12	(0.22)	0.01	(0.10)	0.01	(0.06)	-0.01	(0.02)
avg_aut	0.03	(0.16)	0.01	(0.46)	-0.13	(0.24)	0.04	(0.06)
tot_cost	0.87	(1.83)	0.01	(0.01)	0.14	(1.15)	-0.19	(0.68)
fund_req	-1.25	(2.62)	0.01	(0.01)	-0.21	(1.64)	0.27	(0.97)
fund_req_perc	0.40	(1.83)	-1.7	(2.48)			0.83	(0.65)
yr_num	0.58	(1.38)	0.03	(0.66)	1.49	(1.83)	-0.93	(0.51)
yr_cost	-0.87	(1.83)	0.01	(0.01)	-0.15	(1.15)	0.19	(0.68)
yr_perc	0.22	(1.41)	-1.0	(1.5)	-0.12	(0.70)	0.52	(0.46)
grr_num					0.27	(0.81)	0.58	(0.43)
grr_cost					-0.15	(1.15)	0.19	(0.68)
grr_perc					-0.11	(0.70)	0.24	(0.51)
d1_l1 = PE					0.24	(1.93)	0.78	(0.66)
d1_l1 = SH					0.25	(1.85)	0.59	(0.83)
d1_l2 = LS2	-1.17	(1.88)						
d1_l2 = LS3	-0.72	(2.24)						
d1_l2 = LS4	-0.17	(1.63)						
d1_l2 = LS6	-0.21	(1.70)						
d1_l2 = LS7	0.14	(2.10)						

Table 5: Firth logistic regression models for the datasets of proposals satisfying the antecedent of rules R1-R4. The dependent variable is *peer-review = passed*. Coefficients marked by [*], [**], and [***] are statistically significant at the 90%, 95% and 99% confidence level respectively. Blank cells are due to control variables with unique values (e.g., `d1_lv_1` is always "LS" in rule R1), or to control variables omitted due to high standard errors. *f_partner_num* is not part of the model in order to account for the included-variable bias (see Section 5.1).

than one young researcher to be hired) but the PI has a low productivity record of publications (at most 12) and high uncertainty on the PI's effective contribution (large number of co-authors). The peer-reviewers of the LS area appear to have compensated the lack of knowledge on the skills of the PIs by some prior knowledge or stereotype resorting to the gender of the PI, with females being disadvantaged. This phenomenon is known as *statistical discrimination* or *rational racism* (Romei and Ruggieri, 2013) – as opposed to *taste-based* discrimination which is motivated by prejudice.

Table 5 reports the Firth logistic regression model for the proposals satisfying the antecedent of rule R1. All other factors being equal, female PIs have $e^{-1.64} = 0.194$ the odds of male PIs of passing the peer-review.[9] The coefficient -1.64 is greater (in absolute value) than -0.33, the one computed over the whole dataset of proposals in program P_1 (see Table 3). More important, the coefficient is now significantly non-zero at 90% confidence level.

6.2. Rule R2: physical and analytical chemical sciences in program P_2

A context of possible discrimination for proposals in program P_2 is unveiled by the following rule:

```
R2: (d1_lv2 = PE4) and (tot_cost >= 1,358,000) and
    (age <= 35) => disc=yes
    [prec=1.0] [rec=0.031] [diff=0.194] [OR=4.50]
```

where PE4 is the Physical and Analytical Chemical Sciences panel, at the second level of the ERC hierarchy. The context of rule R2 concerns proposals with high budget led by young PIs. There are 9 proposals led by male PIs, 2 of which passed the peer-review, and 11 proposals led by female PIs, none of which passed it. The recall of rule R2 is 3.1%, i.e., the context covers 3.1% of the proposals labeled as discriminated. The precision of rule R2 is 100%, meaning that all of the 11 proposals in the context led by female PIs have been labeled as discriminated. The average risk difference is 19.4%, and the odds ratio is 4.5.

Table 5 shows the Firth logistic regression model for the proposals in the context of rule R2. All other factors being equal, female PIs have $e^{-1.37} = 0.254$ the odds of male PIs of passing the peer-review. The coefficient -1.37 is greater (in absolute value) than -0.87, the one computed over the whole dataset (see Table 3), yet it is significantly non-zero at the lower confidence level of 95%. Summarizing, rule R2 unveils a niche of proposals with a gender bias higher than the average bias of the whole dataset of proposals in program P_2.

6.3. Rule R3: expensive projects in program P_1

A second rule about proposals in the program P_1 is the following:

```
R3: (yr_cost >= 187,000) and (grr_cost >= 70,000)
     => disc=yes
    [prec=0.86] [rec=0.052] [diff=0.161] [OR=5.77]
```

The antecedent of the rule concerns proposals with high budget for young researchers and for good reputation researchers. We checked that such a context is disjoint from the one of rule R1, where all proposals had no budget for good reputation researchers. There are 16 proposals led by male PIs in the context, 4 of which passed the peer-review,

[9] Here we deal with the odds of *passing*, and low values denote high burden. The odds ratio (OR) deals with the odds of *being rejected*, and high values denote high burden.

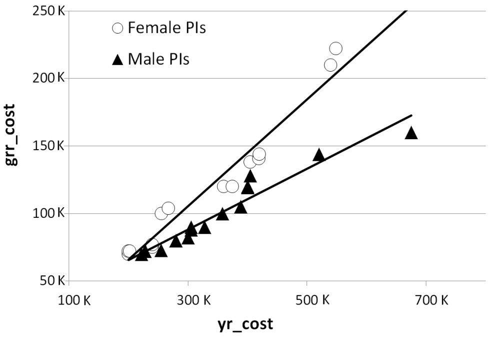


Figure 11: Scatter plot of `grr_cost` over `yr_cost` for proposals satisfying the antecedent of R3.

and 14 proposals led by female PIs, none of which passed it. Precision of the rule is 86%, meaning that 12 out of the 14 proposals led by female PIs have been labeled as discriminated, with an average risk difference for the 14 proposals of 16.1%. The recall of the rule is 5.2%, hence rules R1 and R2 cover together 14.7% of the proposals labeled as discriminated. The odds ratio of the proposals in the context is 5.77.

Intuitively, the peer-reviewers of program P_1 seem to trust more male PIs than female PIs in managing projects with high costs of personnel, namely young and good reputation researchers. Does rule R3 unveil then a case of actual discrimination? Firth logistic regression on the dataset of proposals of the context (see Table 5) shows a coefficient for *gender=Female* of -0.86, which, however, is not statistically significant at 90% confidence level – i.e., the null hypothesis that the coefficient is actually zero cannot be rejected. We proceed by analysing further the proposals in the context. The scatter plot in Figure 11 reports the costs of good reputation researchers over the costs of young researchers. It highlights that proposals led by female PIs tend to a higher proportion of good reputation researcher costs over proposals led by male PIs. This could somehow be in contrast with the intended objectives of the call for proposals, which require a substantial hiring of young researchers. Therefore, it may well be the case that peer-reviewers have scored worse those proposals relying too much on the hiring of senior researchers. This could be argued as a legitimate and objective justification for the disparate treatment of female PIs, an exception admitted by the anti-discrimination laws. Whether this is the case or not, however, is a matter of legal argumentation. Strictly speaking, the call for proposals did not set an explicit maximum threshold on the proportion of good reputation researcher costs over young researcher costs.

6.4. Rule R4: young PIs in program P_2

A second rule for proposals in the program P_2 is the following:

```
R4: (age <= 32) and (fund_req >= 310,000)
    => disc=yes
    [prec=0.52] [rec=0.12] [diff=0.07] [OR=9.6]
```

The antecedent of the rule concerns younger PIs with a fund request greater or equal than € 310K. Intuitively, this can be interpreted as a negative bias against younger female PIs who require a medium-high grant. There are 201 proposals in such a context: 131 with male PIs, 16 of which passed the peer-review; and 70 with female PIs, only 1 of which passed the peer-review. The odds ratio is 9.6. Precision of the rule is moderately higher than 38% – the overall percentage of proposals labeled as discriminated in program P_2. That is, about half of the 69 female PIs whose proposal was rejected showed a risk difference greater or equal than 10%. In fact, the average risk difference is only 7%. However, recall is rather high: 12% of the proposals labeled as discriminated in program P_2 are in the context of rule R4. Finally, we checked that the overlap of proposals in the contexts of both rule R4 and R2 is minimal, with only 3 proposals led by male PIs and 1 led by a female PI. Such overlap originates from the fact that rules R2 and R4 are selected from two different classification models.

Consider the logistic regression model for proposals in the context of rule R4 (see Table 5). All other factors being equal, female PIs have $e^{-1.20} = 0.301$ the odds of male PIs of passing the peer-review. The coefficient -1.20 is smaller (in absolute value) than the one of rule R2, but significant at the higher confidence level of 99%. Moreover, it is greater (in absolute value) than the one of the whole dataset (see Table 3). Summarizing, rule R4 highlights a context with higher gender bias than in the whole dataset of proposals in program P_2. This and the contexts of the other rules were not previously known as possible stereotypes of discriminatory behaviors. Rather, they have been the result of a discrimination discovery investigation.

7. A KDD process in support of discrimination discovery

Since personal data in decision records are highly dimensional, i.e., characterized by many multi-valued variables, a huge number of possible contexts may, or may not, be the theater for discrimination. In order to extract, select, and rank those that represent actual discriminatory behaviors, an anti-discrimination analyst should apply appropriate tools for pre-processing data, extracting prospective discrimination contexts, exploring in details the data related to the context, and validating them both statistically and from a legal perspective[10]. Discrimination discovery consists then of an iterative and interactive process. Iterative because, at certain stages, the user should have the possibility of choosing different algorithms, parameters, and evaluation measures or to iteratively repeat some steps to unveil meaningful discrimination patterns.

[10] As observed by Gastwirth (1992), the objectives of science and the law often diverge, with rigorous scientific methods conflicting with the adversarial nature of the legal system.

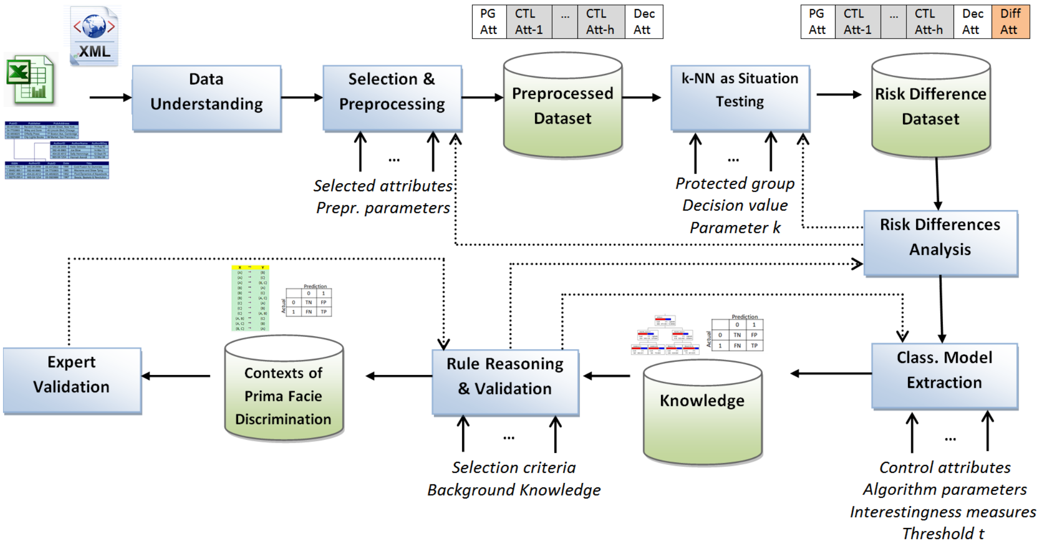


Figure 12: The KDD process of situation testing for discrimination discovery.

Interactive because several stages need the support of a domain expert in making decisions or in analysing the results of a previous step. We propose here to adopt the process reported in Figure 12, which is specialized in the use of the situation testing for extracting contexts of possible discrimination. The process has been abstracted from the case study presented in the previous sections, and it consists of four major steps.

Data Understanding and Preparation. The availability of historical data concerning decisions made in socially-sensitive tasks is the starting point for discovering discrimination. We assume a collection of data sources storing historical decisions records in any format, including relational, XML, text, spreadsheets or any combination of them. Standard data pre-processing techniques (selection, cleansing, transformation, outlier detection) can be adopted to reach a pre-processed dataset consisting of an *input relation* as the basis for the discrimination analysis. The grain of tuples in the relation is that of an individual (an applicant to a loan, to a position, to a benefit). Three groups of attributes are assumed to be part of the relation:

protected group attributes: one or more attributes that identify the membership of an individual to a protected group. Attributes such as sex, age, marital status, language, disability, and membership to political parties or unions are typically recorded in application forms, curricula, or registry databases. Attributes such as race, skin color, and religion may be not available, and must be collected, e.g., by surveying the involved people;

decision attribute: an attribute storing the decision for each individual. Decision values can be nominal, e.g., granting or denying a benefit, or continuous, e.g., the interest rate of a loan or the wage of a worker;

control attributes: one or more attributes on control factors that may be (legally) plausible reasons that may affect the actual decision. Examples include attributes on the financial capability to repay a loan, or on the productivity of an applicant worker.

Risk Difference Analysis. For each tuple of the input relation denoting an individual of the protected group, the additional attribute *diff* is calculated as the risk difference between the decisions of its k nearest-neighbors of the protected group and the decisions for its k nearest-neighbors of the unprotected group (see Section 2.3). We call the output of the algorithm the *risk difference relation*. The value k is a parameter of the algorithm. A legitimate question is how to choose the "right" k? A large k means that every instance is a neighbor, hence the distribution of *diff* tends towards a unique value. Conversely, for a small k, we run the risk that the distribution is affected by randomness. As a consequence, a study of the distribution of *diff* for a few values of k is required. This means iterating the calculation of the *diff* attribute. Exploratory analysis of *diff* distributions may also be conducted to evaluate risk differences at the variation of: the protected group under consideration, e.g., discrimination against women or against youngsters; the compound effects of multiple discrimination grounds, e.g., discrimination against young women vs discrimination against women or youngsters in isolation; the presence of favoritism towards individuals of a dominant group, e.g., nepotism. Once again, this requires iterating the calculation of *diff* by specifying a different protected group attribute to focus on.

Discrimination Model Extraction. By fixing a threshold value t, an individual $\mathbf{r}$ of the protected group can be labeled as discriminated or not on the basis of the condition $\mathit{diff}(\mathbf{r}) \geq t$. We introduce a new boolean attribute *disc* and set it to true for a tuple $\mathbf{r}$ meeting the condition above, and to false otherwise. A global description of who has been discriminated can now be extracted by resorting to a standard classification problem on the dataset of individuals of the protected group, where the class attribute is the newly introduced *disc* attribute. Accuracy of the classifier is evaluated with objective interestingness measures, e.g., precision and recall over the *disc* = *true* class value. The intended use of the classifier is descriptive, namely to provide the analyst with a characterization of the individuals that have been discriminated. The choice of the value t should then be supported by laws or regulators.[11] For instance, the *four-fifths rule* by the US Equal Employment Opportunity Commission (1978) states that a job selection rate (the RC measure from Figure 1) lower than 80% represents a *prima facie* evidence of adverse impact.

Since the intended use of the extracted classifier is descriptive, classification models that are easily interpretable by (legal) experts and whose size is small should be preferred. In other words, one should trade accuracy for simplicity. Classification rules and decision trees are natural choices in this sense, since rules and tree paths can easily be interpreted and ranked. The extracted classification models provide a global description of the *disc* class values. They are stored in a knowledge base, for comparison purposes and for the filtering of specific contexts of discrimination – as described next.

Rule Reasoning and Validation. The actual discovery of discriminatory situations and practices may reveal itself as an extremely difficult task. Due to time and cost constraints, an anti-discrimination analyst needs to put under investigation a limited number of contexts of possible discrimination. In this sense, only a small portion of the classification models can be analysed in detail, say the top N rules or the top N paths of a decision tree. We propose to concentrate on classification rules of the form:

```
(cond_1) and ... and (cond_n) =>
              disc=yes [prec] [rec] [diff] [OR]
```

where `(cond_1) and ... and (cond_n)` is obtained from a classification model (from a rule or from a path of a decision tree). Rules are ranked on the basis of one or more interestingness measures, including: precision `[prec]` (proportion of discriminated individuals among those of the protected group which satisfy the antecedent), recall `[rec]` (proportion of the overall discriminated individuals covered by the antecedent), average value of *diff* `[diff]` (a measure of the degree of discrimination observed by individuals of the protected group which satisfy the antecedent), and odds ratio `[OR]` (a measure of the burden of negative decisions on the individuals of the protected group when compared to those of the unprotected group satisfying the antecedent of the rule). Notice that `[diff]` and `[OR]` may rank rules differently because they contrast distinct sets of groups (the $2k$ nearest neighbors, and the members of the unprotected group satisfying the antecedent of the rule). Statistical validation is accounted for in our approach by relying on logistic regression, which is a well-known tool in the legal and economic research communities. Readability and interpretability should also be taken into account by preferring rules with fewer items in the antecedent, thus trading interestingness with simplicity.

As an alternative approach to the selection of rules from the classifiers extracted in the previous step of the process, one could mine *all* classification rules of the form above by means of association rule mining. Unfortunately, this results in a huge number of rules covering overlapping contexts of possible discrimination. This is what occurs, for instance, in the rule extraction and filtering approach of Ruggieri et al. (2010a). The rules selected in this step of the process, however, are still subject to further consideration, e.g., by a legal expert, who may require further data exploration and, possibly, iteration of previous steps of the process. Therefore, the number of selected rules must be reasonably low. Selecting the best rules from the best performing classification models is then a means to keep the number of (overlapping) rules to a minimum.

8. Conclusions

The contribution of this paper has been threefold.

First, we have presented a complex case study in the context of scientific project funding using real data from an Italian national call for proposals. The application of discrimination discovery methodologies based on data mining to real case studies was lacking in the existing literature. So far, experiments and analyses have been conducted on "general purpose" datasets, not explicitly collected or processed for discrimination analysis. As a consequence, the reported analyses have been necessarily partial, typically being limited to summary statistics (e.g., number of possibly discriminatory contexts found), to artificial examples, and to generic argumentations on the results found. This is a serious drawback that limits the acceptance of knowledge discovery methods in practice.

Second, we have proposed and applied a methodology that couples legal methods (situation testing) for the definition of cases of possible discrimination, data mining methods (a variant of k-NN plus standard classification) for the search of contexts of possible discrimination, and regression analysis for the statistical validation of such contexts. This approach overcomes the statistical analysis[12] of discrimination conducted in the social sciences, economics, and legal literature, which is limited to the verification of an hypothesis of possible discrimination on

[11] A relevant question is the other way round – namely, can data mining help law makers and regulators in the definition of appropriate values for t?

[12] The contrast between the two approaches above is an instance of the two general "cultures" in the use of statistical modeling (Breiman, 2001): data modeling vs the algorithmic modeling.

the *whole* set of past decision records. Such an analysis reveals to be inadequate to cope with the problem of *searching for* unknown or unforeseen contexts of discriminatory decisions hidden in a large dataset. On the contrary, the rules discussed in Section 6 unveil *prima facie* evidence of discrimination when certain project costs are above a threshold value. Both the cost attribute and the threshold value, however, come as the result of the analysis – they were not an a priori hypothesis to be verified. The extraction of contexts of discrimination is precisely the objective of discrimination discovery.

Third, from the specific case study, we have abstracted a general process of discrimination discovery. The adopted methodology relies on an implementation of the legal practice of situation testing using a variant of k-NN, and then on extracting and reasoning about a classification model. The steps of the methodology have been described in the process of Figure 12, which represents a guidance for researchers and anti-discrimination analysts. We believe that this contribution can provide higher confidence about the replicability of the analyses and their applicability to real cases.

Some issues remain open for future investigation. With reference to the case study, further analysis will be made possible by enriching the available dataset with additional control features, e.g., some accurate measures of the scientific productivity of applicants and of their professional network. This was not possible in our study, since our input data were anonymized. Concerning the tools adopted, while the k-NN algorithm remains the core component of the proposed process, of particular interest is the formalization of the deductive component, in which the extracted classification models are filtered, refined, transformed and validated into useful knowledge. We aim at designing a post-processing tool, by adapting the XQuake system (Romei and Turini, 2010), able to support the user in the deductive part of the process. Finally, throughout the paper, we have assumed the availability of a feature denoting the protected group under analysis – in the case study, the gender of the PIs. In indirect discrimination discovery, this assumption does not hold, e.g., because race, ethnicity, or sexual orientation may be not recorded in data. In such cases, a different approach must be devised.

References

Bendick, M., 2007. Situation testing for employment discrimination in the United States of America. Horizons Stratégiques 3 (5), 17–39.

Bentley, J. T., Adamson, R., 2003. Gender differences in the careers of academic scientists and engineers: A literature review. Special report, National Science Foundation, `http://www.nsf.gov`.

Bornmann, L., Daniel, H.-D., 2005. Selection of research fellowship recipients by committee peer review: Reliability, fairness and predictive validity of Board of Trustees' decisions. Scientometrics 63 (2), 297–320.

Bornmann, L., Daniel, H.-D., 2009. The state of h index research. EMBO reports 10 (1), 2–6.

Bornmann, L., Mutz, R., Daniel, H.-D., 2008. Latent markov modeling applied to grant peer review. Journal of Informetrics 2 (3), 217–228.

Breiman, L., 2001. Statistical modeling: The two cultures. Statistical Science 16 (3), 199–231.

Brouns, M., 2000. The gendered nature of assessment procedures in scientific research funding: The Dutch case. Higher Education in Europe 25 (2), 193–199.

Calders, T., Verwer, S., 2010. Three naive bayes approaches for discrimination-free classification. Data Mining & Knowledge Discovery 21 (2), 277–292.

Ceci, S. J., Williams, W. M., 2011. Understanding current causes of women's underrepresentation in science. Proc. of the National Academy of Sciences 108 (8), 3157–3162.

Cohen, W. W., 1995. Fast effective rule induction. In: Proc. of Int. Conf. on Machine Learning (ICML 1998). Morgan Kaufmann, pp. 115–123.

Council of the E.U., 1999. Resolution 1999/C 201/01 on Women and Science. `http://eur-lex.europa.eu`.

Custers, B. H. M., Calders, T., Schermer, B. W., Zarsky, T. Z. (Eds.), 2013. Discrimination and Privacy in the Information Society. Vol. 3 of Studies in Applied Philosophy, Epistemology and Rational Ethics. Springer.

Equal Employment Opportunity Commission, 1978. Uniform guidelines on employee selection procedure. 43 FR 38295, `http://www.gpo.gov`.

European Commission, 2009. The gender challenge in research funding: Assessing the European national scenes. Directorate General for Research, Science, Economy and Society, Unit L.4, `http://ec.europa.eu`.

European Commission, 2012. Meta-analysis of gender and science research. Directorate General for Research and Innovation, Sector B6.2, `http://www.genderandscience.org`.

Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80 (1), 27–38.

Frank, E., Witten, I. H., 1998. Generating accurate rule sets without global optimization. In: Proc. of Int. Conf. on Machine Learning (ICML 1998). Morgan Kaufmann, pp. 144–151.

Gastwirth, J. L., 1992. Statistical reasoning in the legal setting. The American Statistician 46 (1), 55–69.

Goldstein, H., 2011. Multilevel Statistical Models, 4th Edition. Wiley.

Hajian, S., Domingo-Ferrer, J., 2012. A methodology for direct and indirect discrimination prevention in data mining. IEEE Transactions on Knowledge and Data Engineering, to appear.

Heinze, G., 2006. A comparative investigation of methods for logistic regression with separated or nearly separated data. Statistics in Medicine 25 (24), 4216–4226.

Jayasinghe, U. W., Marsh, H. W., Bond, N. W., 2003. A multilevel cross-classified modeling approach to peer-review of grant proposals. Journal of the Royal Statistical Society 166 (3), 279–300.

Kamiran, F., Calders, T., 2012. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems 33, 1–33.

Killingsworth, M. R., 1993. Analyzing employment discrimination: From the seminar room to the courtroom. American Economic Review 83 (2), 67–72.

Larivière, V., Vignola-Gagné, E., Villeneuve, C., Gélinas, P., Gingras, Y., 2011. Sex differences in research funding, productivity and impact: an analysis of Québec university professors. Scientometrics 87 (3), 483–498.

Ley, T. J., Hamilton, B. H., 2008. The gender gap in NIH grant applications. Science 322 (5907), 1472–1474.

Luong, B. T., Ruggieri, S., Turini, F., 2011. k-NN as an implementation of situation testing for discrimination discovery and prevention. In: Apté, C., Ghosh, J., Smyth, P. (Eds.), Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2011). ACM, pp. 502–510.

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.

Mutz, R., Bornmann, L., Daniel, H.-D., 2012. Does gender matter in grant peer review? An empirical investigation using the example of the Austrian Science Fund. Journal of Psychology 220, 121–129.

Pager, D., 2007. The use of field experiments for studies of employment discrimination: Contributions, critiques, and directions for the future. The ANNALS of the American Academy of Political and Social Science 609 (1), 104–133.

Pedreschi, D., Ruggieri, S., Turini, F., 2008. Discrimination-aware data mining. In: Li, Y., Liu, B., Sarawagi, S. (Eds.), Proc. of the ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD 2008). ACM, pp. 560–568.

Quillian, L., 2006. New approaches to understanding racial prejudice and discrimination. Annual Review of Sociology 32 (1), 299–328.

Quinlan, J. R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA.

RAND, 2005. Is there gender bias in federal grant programs? RAND Infrastructure, Safety, and Environment Brief RB-9147-NSF, http://rand.org.

Romei, A., Ruggieri, S., 2013. Discrimination data analysis: A multidisciplinary bibliography. In: Custers, B. H. M., Calders, T., Schermer, B. W., Zarsky, T. Z. (Eds.), Discrimination and Privacy in the Information Society. Vol. 3 of Studies in Applied Philosophy, Epistemology and Rational Ethics. Springer, pp. 109–135.

Romei, A., Ruggieri, S., Turini, F., 2012. Discovering gender discrimination in project funding. In: Proc. of the IEEE ICDM 2012 Int. Workshop on Discrimination and Privacy-Aware Data Mining (DPADM). IEEE Computer Society, pp. 394–401.

Romei, A., Turini, F., 2010. XML Data Mining. Software: Practice and Experience 40 (2), 101–130.

Rorive, I., 2009. Proving Discrimination Cases - the Role of Situation Testing. Centre For Equal Rights & Migration Policy Group, http://www.migpolgroup.com.

Ruggieri, S., Pedreschi, D., Turini, F., 2010a. Data mining for discrimination discovery. ACM Trans. on Knowledge Discovery from Data 4 (2), Article 9.

Ruggieri, S., Pedreschi, D., Turini, F., 2010b. DCUBE: Discrimination discovery in databases. In: Elmagarmid, A. K., Agrawal, D. (Eds.), Proc. of the ACM SIGMOD Int. Conf. on Management of Data (SIGMOD 2010). ACM, pp. 1127–1130.

Sandström, U., Hällsten, M., 2008. Persistent nepotism in peer-review. Scientometrics 74 (2), 175–189.

UNESCO, 2007. Science, Technology and Gender: An International Report, 4th Edition. UNESCO Publishing.

Wennerås, C., Wold, A., 1997. Nepotism and sexism in peer-review. Nature 387 (5), 341–343.

Wilson, R., 2004. Where the elite teach, it's still a man's world. The Chronicle of Higher Education 51 (15).

Witten, I. H., Frank, E., 2011. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations., 3rd Edition. Morgan Kaufmann, San Francisco.