Gender Disparities in Child Custody Sentencing in Spain: a Data Driven Analysis

Júlia Riera
Universitat Pompeu Fabra
Barcelona, Spain
julia.riera.perramon@gmail.com

David Solans
Telefónica Research
Barcelona, Spain
david.solansnoguero@telefonica.com

Marzieh Karimi-Haghighi
Universitat Pompeu Fabra
Barcelona, Spain
m.karimihaghighi@gmail.com

Carlos Castillo
ICREA and Universitat Pompeu Fabra
Barcelona, Spain
chato@icrea.cat

Caterina Calsamiglia
ICREA at IPEG
Barcelona, Spain
caterina.calsamiglia@barcelona-ipeg.eu

ABSTRACT

In our work, we investigate biases in judicial decisions using data analytics. Specifically, we are interested in analyzing the impact of the gender of both the judge and the plaintiff on the probability of winning a case. With this aim, we analyze a dataset comprising information from over one thousand second-instance appeals for child custody in Spain. Our results indicate significant differences in how legal arguments and facts are utilized in the final sentences, depending on the gender of the plaintiff. We also examine the impact of the requested type of custody (sole or joint) on the probability of winning a case, with a focus on its relationship to the plaintiff's gender. Moreover, our analysis reveals statistically significant differences in the winning probability of a case depending on the judge's gender. To further understand these findings, we conduct additional analysis to establish the causal relationship between the judge's gender and the probability of winning, revealing weak but consistent patterns. Our research provides a consistent methodology for evaluating biases in judicial systems and offers intriguing insights into the context of child custody in Spain.

CCS CONCEPTS

• **Mathematics of computing** → **Exploratory data analysis**; • **Applied computing** → **Law**.

KEYWORDS

implicit bias, child custody, judicial system, data analytics

ACM Reference Format:
Júlia Riera, David Solans, Marzieh Karimi-Haghighi, Carlos Castillo, and Caterina Calsamiglia. 2023. Gender Disparities in Child Custody Sentencing in Spain: a Data Driven Analysis. In *Nineteenth International Conference on Artificial Intelligence and Law (ICAIL 2023), June 19–23, 2023, Braga, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3594536.3595135

ICAIL 2023, June 19–23, 2023, Braga, Portugal

ACM ISBN 979-8-4007-0197-9/23/06…$15.00
https://doi.org/10.1145/3594536.3595135

1 INTRODUCTION

The term *"implicit bias"* refers to attitudes or stereotypes that unconsciously influence our understanding, decision-making, and behavior [16]. In this study, we conduct a data-driven analysis to quantify the potential effects of implicit biases related to gender in the context of child custody sentencing in Spain.

Recent data from the Spanish National Institute of Statistics (INE) [1] shows that the number of divorces in Spain has been increasing year after year. In 2021, there were 86,851 divorces, representing a 12.5% increase from 2020. Furthermore, 21.5% of these cases were contentious divorces requiring judicial consideration in courts of first instance. Among these cases, 56.8% involved marriages with underage or economically dependent children, requiring a determination of child custody type. This means that just in 2021, more than 10,000 child custody sentences were issued in Spain, highlighting the importance of identifying potential gender biases in this domain.

Although good intentions and objectivity are assumed in judicial decisions, unintended effects of implicit biases may still influence judicial outcomes. An example of the harm unintended biases may cause in the judicial system is the *"the hungry judge effect"*, which suggests that judges are more likely to make favorable rulings before breaks [8]. Effects like this can lead to significant inequalities at the individual level, causing similar cases to receive different sentencing. However, the presence of gender bias can be more dangerous due to its nature as a group-based bias, leading to systematic loss of opportunities for individuals of the specific gender that is given unprivileged treatment.

We argue that, although it can be challenging, raising awareness is a crucial initial step towards mitigating the effects of implicit biases. To contribute to this effort, we study gender biases present in historical records and analyze an annotated dataset containing more than 1,800 second-instance sentences in the context of child custody in Spain. We consider the gender of the plaintiff and the judge, as well as additional factors such as the type of custody, family home allocation, and maintenance payment allocation. Additionally, we examine the legal principles and facts highlighted in each sentence.

Machine learning (ML) systems are increasingly being adopted in the judicial system, with applications ranging from risk prediction

[1] INE, Divorce statistics in Spain. https://www.ine.es/prensa/ensd_2021.pdf

to the detection of borderline cases [15, 17, 1]. In ML, data is often referred to as "an initiator of biases" that reflects social stereotypes and prejudices [5, 13]. Previous research in the context of the judicial system has emphasized the importance of addressing fairness before deploying ML systems [1]. Our work studies this potential input data with the aim of describing the child custody resolution process through gender lenses to get a deeper understanding of how gender affects it. That way, possible biases present in the data can be taken into account or mitigated before training subsequent ML systems on the analyzed dataset [22].

The analyzed dataset contains second instance judgments, which refer to cases in which a court re-hears and delivers a judgment on a case that was originally heard in a court of first instance. In these cases, the plaintiff of the second instance appeals the judgment of the first instance. In order to quantify gender biases, we take into consideration the gender of both the plaintiff and the judge. In our experiments, we treat gender as binary, as it is represented in the analyzed dataset. In addition to the gender of the judge and plaintiff, the dataset contains information about the plaintiff's request in terms of (i) type of custody (sole custody vs joint custody); (ii) family home allocation (suppression vs attribution), and; (iii) maintenance payment allocation (suppression vs attribution). Additionally, it contains an annotated collection of legal principles and facts highlighted by each of the sentences.

Our research presents a robust methodology for assessing biases within judicial systems. The approach involves an extensive exploratory data analysis, with a special focus on gender, which can be found in Section 3. Subsequently, in Section 4 we investigate the direct effect of gender through feature importance analysis to detect any blatant gender bias. Furthermore, we perform an indirect effect analysis in Section 5, utilizing methods to identify other ways in which gender can influence the judicial outcome through other variables. In addition, in Section 7 we incorporate an analysis that employs counterfactual fairness to identify how gender can impact similar sentences. Overall, this methodology offers a comprehensive framework for evaluating biases in judicial systems, and we believe it will be useful for future research in this field.

2 RELATED WORK

2.1 Implicit bias in judicial decisions

The role of implicit biases in judicial decisions has been widely studied in the literature. Among the extensive body of research that has been conducted in the area, we begin by highlighting two highly influential works in the broad domain of judicial decisions, and then focus on the field of child custody and family law cases.

In 2011, researchers highlighted the "hungry judge effect" [8], which revealed that judges were more inclined to be lenient after a meal but more severe before the break. This suggests that the scheduling of cases might have a direct effect on their outcome and duration.

One year later, Kang et al. [16] provided an extensive introduction to implicit bias and analyzed two different real cases, evidencing the presence of implicit biases that had an effect on legal outcomes. They proposed various intervention strategies to mitigate implicit bias and limit its effect on behavior. Among the strategies they proposed is the need to question one's objectivity and improve decision-making conditions. Our work is highly related to this strategy, as identifying systematic biases in judicial decisions might help judges question their objectivity when applying legislation.

Focusing on the context of family law cases, Costa et al. [7] studied the role of gender stereotypes in child custody decisions in Argentina, Brazil, and the United States. Among other interesting findings, their work points to asymmetries in custody awards, driven by the tendency to ascribe to mothers traits such as friendliness, generosity, or trustworthiness, which are often associated with females over males. This tendency caused higher custody awards for female applicants in their analyzed dataset.

Additionally, more recent findings [18] show that "Pro Se litigants" - those who are unrepresented by legal counsel - are disadvantaged by the presence of cognitive biases in legal officials - judges and attorney-mediators - in the court of family law cases. Considering that whether parties are represented or not in family law cases largely depends on whether parties have the ability to pay for a lawyer, the presence of such cognitive biases implies systematic disadvantageous treatment against individuals with lower economic capacities. This corresponds to another example of how unintended biases can lead to systematic biases against demographic groups in the same domain where our study was conducted.

Autonen-Vaaraniemi [3] performed a qualitative analysis to examine divorce professionals' attitudes and stances in response to common criticisms of how they deal with divorce outcomes for fathers. The interviewed professionals agreed on the need of treating both parents equally but at the same time strove to position themselves as gender-neutral and promoters of equality between mothers and fathers. These findings suggests the presence of systematic biased attitudes against fathers between different types of professionals working in the field.

Our study aims to contribute to the understanding of potential biases in the context of child custody sentencing in Spain by analyzing a dataset of second-instance sentences, and examining the gender of the plaintiff and judge, as well as additional factors such as the type of custody, family home allocation, and maintenance payment allocation. Additionally, we have examined the legal principles and facts highlighted in each sentence.

The results of our study have the potential to inform interventions and policies aimed at reducing the influence of implicit biases in the judicial system and promoting fairness in child custody sentencing. Furthermore, by identifying potential biases in the input data, our study can also inform the development and deployment of machine learning systems in the legal domain, with the goal of ensuring fairness and reducing bias.

In conclusion, our study adds to the growing body of literature on the impact of implicit biases in the judicial system and highlights the importance of considering the role of gender in child custody sentencing. The results obtained in this study may not be generalizable to other jurisdictions, cultures, or domains, but it can serve as a starting point for further research on the topic and to raise awareness about the presence of bias in the legal system.

2.2 Data analytics for quantifying biases in justice system

The use of data-driven analysis in judicial systems has been gaining increasing attention as a means of identifying unintended biases. With the growing adoption of machine learning (ML) systems in the field, data-driven analytics can play a crucial role in building "fair" and "bias-free" ML-based systems.

Previous research has highlighted the importance of addressing fairness and bias issues in ML-based legal systems. For example, Ashley K.[1] reviewed the use of ML in the field of law and emphasized the need to address these issues to ensure the applicability of these systems. Similarly, Bex et al.[4] reviewed ML algorithms used to predict legal outcomes and emphasized the importance of transparency in understanding and mitigating potential biases in the predictions.

In the criminal justice system, Karimi-Haghihi et al.[17] proposed a methodology to improve algorithmic fairness in ML predictions of recidivism risk.

Additionally, studies in the field of family law, such as Costa et al.[7], have highlighted the presence of gender stereotypes and asymmetries in child custody decisions.

Data-driven analysis is a powerful tool for identifying and understanding biases in the judicial system, particularly in child custody sentencing. It involves using data, statistical methods, and machine learning techniques to identify patterns and relationships that may indicate the presence of bias. As other works in the general domain of judicial decisions have done previously, our study uses data-driven analysis to quantify the influence of gender-based implicit biases in child custody sentencing in Spain.

3 DATASET AND DATA PRE-PROCESSING

The data for this study was collected by the Bidaraciv project, conducted by Universidad de Zaragoza and Aragón's Technological Institute ITAINNOVA [9] [10].

3.1 Overview

The dataset used encompasses 1884 second-instance child custody rulings issued in Spain between 2015 and 2020 [2]. These records provide valuable insights into the legal system's handling of child custody matters during this time period. It's important to consider that the analyzed dataset is limited to second-instance sentences, which should be kept in mind when drawing general interpretations from the results in this study. Each record in the dataset provides information about the plaintiff's requests, arguments used by the court, and court decisions. It also includes gender information for the judge, plaintiff, and defendant.

Between the set of available features related to the plaintiff's request, there is information about the type of custody requested by the plaintiff together with economic features related to maintenance payments and family home. The type of custody is represented by a binary attribute with values 0 (sole custody) or 1 (joint custody). Regarding maintenance payments, the plaintiff can request for them to be established or for their amount to be increased (referred to as "attribution"), or for them to be suppressed or decreased (referred to as "suppression"). For the family home, the plaintiff can ask for its attribution, or its suppression.

The court arguments are mainly categorized as legal principles (LP) or facts (FC). The ontology applied to label the data consists of four legal principles and eight facts, as listed in Table 3.

The court decision variables are related to the resolution of the plaintiff's request. A "won" sentence is one where the requested and resolved custody types are the same. In this study, the "victory condition" is used as the target variable.

3.2 Data pre-processing

We employ web scraping to retrieve unique and anonymous IDs for the judges involved in each sentence, as well as the regions of Spain where the sentences were issued. Based on that, we identify 314 unique judge IDs with a 60% prevalence of males. Additionally, we observe a non-uniform distribution of cases across regions, with the highest number of cases (321) in Catalonia and the lowest (3) in Melilla. The mean number of sentences per region is 94.2, with a standard deviation of 91.37. The availability of unique judge IDs enables performing analysis at a judge level, comparing sentences across various judges or for each of them individually.

Sentences where the plaintiff and defendant have the same gender are excluded from further analysis due to the small sample size (only 3 out of 1884 records).

The dataset is normalized using Min-Max scaling, to scale all features that are neither binary nor categorical to a range of [0,1].

3.3 Descriptive statistics

We conduct an exploratory analysis that evaluates the base rates between gender variables in the dataset. Specifically, we analyze variables such as court decision, request type, and victory condition in relation to the plaintiff and judge genders. After identifying some disparities, we perform hypothesis testing to determine the statistical significance of these, using a two-sample proportion z-test.

3.3.1 Plaintiff. Male comprise 60.03% of the plaintiffs in the dataset. When examining court decisions, 82.22% of the cases have a decision that is opposite to the plaintiff's custody type request, while only 17.78% have a decision that is in favor of the plaintiff's custody type request. The dataset includes 54.81% joint custody requests and 45.19% sole custody requests. Sole custody requests have a victory rate of 13.48%, while joint custody requests have a victory rate of 21.34%. This represents a relative difference of 58% in favor of joint custody requests.

Furthermore, our analysis shows that female plaintiffs tend to request sole custody at a higher rate (86.45%) compared to males (17.86%) whereas male plaintiffs are more likely to request joint custody (82.14%) compared to females (13.55%).

In terms of winning condition, where the requested and granted custody type match, females win 14.34% of the cases, while males win 20.07%. However, we hypothesize that these differences are a result of males being more likely to request joint custody, which has a higher success rate, as previously mentioned. This hypothesis is supported by the results of the two-sample proportion z-test, which shows that while there is a significant difference when looking at

[2]Example of a sentence: http://labje.unizar.es/sentencias/APA_2018_3010.pdf

all the data (p-value of 0.001), this difference is not present when analyzing data by request type (p-values of 0.31 and 0.14 for sole and joint custody, respectively). This suggests that the observed differences are driven by the disparate request types made by each gender group.

Additionally, we observe that females have relatively similar chances of winning regardless of whether they request sole or joint custody (14.13% and 15.69% respectively). However, for males, the difference in winning rates between sole and joint custody is greater (11.39% and 21.96% respectively). It's important to note that, despite males having higher overall victory rates, they have lower chances of winning sole custody compared to females, but higher probabilities of winning when requesting joint custody.

3.3.2 Judge. In the dataset, 63.38% of judges are male and 36.62% are female. There are 314 unique judge IDs. Out of these, 59.24% are male and 40.76% are female. The assignation of sole and joint custody sentences among male and female judges is evenly balanced, with both genders receiving joint custody cases slightly more often. Females are assigned 45.07% of sole custody cases and 54.93% of joint custody cases, while males are assigned 45.39% and 54.61%, respectively.

The victory rate, or the rate at which the requested custody is granted, reveals a 28% difference between female and male judges. The percentage of won cases among the ones seen by female and male judges is 20.58% and 16.16% respectively. When dividing the data by custody type, for sole custody, victory rates are 16.72% and 11.62% for female and male judges respectively. For joint custody, victory rates are 23.75% and 19.94% for female and male judges respectively. The two-sample proportion z-test shows a significant difference in the victory rates among genders (p-value=0.015). However, this significant difference only holds when analyzing all data or data for sole custody cases. No significant difference is found in joint custody cases.

Continuing our analysis of the victory rates, we tested the hypothesis that judges tend to favor plaintiffs of the same gender [6]. Our findings show that a female judge increases the chances of winning for both male and female plaintiffs, and male plaintiffs are more likely to win regardless of the judge's gender.

The higher victory rates for male plaintiffs and female judges prompted further investigation to determine the extent to which these differences can be attributed to biases.

3.4 Economic features

There are two economic features related to the plaintiff's request: maintenance payments and family home. These features give insight into the economic state of the plaintiff, with differences between the two genders potentially indicating disparities in economic needs or economic dependence. Parts of the analysis are also performed on data separated by request type, given our observation that economic features are highly tied to the type of custody requested. For instance, sole custody requests are more likely to also ask for the attribution of maintenance payments (39.04%) compared to joint custody requests (3.01%).

Our analysis of economic features shows an imbalanced ratio between requests for attribution and suppression of resources across genders. Females are more likely to request attribution while males are more likely to request suppression. This pattern holds true when dividing the data by custody type in the case of maintenance payments, as seen in Table 1. The same pattern is seen for family home requests in sole custody cases, but not in joint custody cases (see Table 2).

Overall, females more often ask for economic support in the form of maintenance payment and/or family home attribution and males are more likely to request the suppression of these resources, suggesting that they were allocated more frequently to females in the first instance judgements.

Table 1: Maintenance payments request percentage by plaintiff gender

	Sole custody		Joint custody	
	Attr.	Supp.	Attr.	Suppr.
Female plaintiff	41.94%	8.76%	6.86%	41.18%
Male plaintiff	29.7%	31.19%	2.58%	58.23%

Table 2: Family home request percentage by plaintiff gender

	Sole custody		Joint custody	
	Attr.	Supp.	Attr.	Suppr.
Female plaintiff	12.29%	0.61%	6.86%	4.9%
Male plaintiff	7.43%	1.98%	8.29%	6.57%

Figure 1 shows minimal variation in the grant rate of requests. In most cases, granted and not granted requests are evenly distributed, with roughly equal chances of approval. However, there are exceptions, such as asking for the suppression of maintenance payments or for the attribution of the family home, where the denial rate is higher. Nonetheless, it is higher for both gender groups.

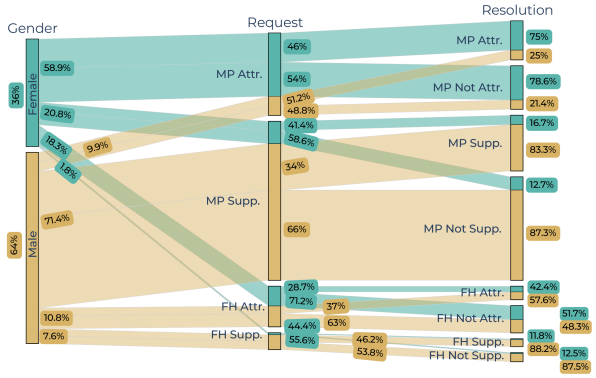
Despite the balanced grant rate, the rightmost part of Figure 1 reveals an imbalance in the total percentages of granted attribution or suppression across genders, which is caused by the unequal distribution of requests between genders.

Females requesting and receiving more economic support clearly reflects female economic dependence. This is likely due to the ongoing gender wage gap in the country [20], where males earn an average of 24.22% more than females[3].

Furthermore, divorced females are a particularly vulnerable group, having potentially lost independence and power during their marriage and while raising children. In many cases, in order to care for the children and due to the lack of paternal co-responsibility, females are compelled to reduce their working hours or forego professional development. Meanwhile, males frequently continue to advance professionally and attain higher salaries. As a result, there is a disparity in economic circumstances between the genders in divorce cases [14].

[3]INE, Average yearly wage across genders in Spain. https://www.ine.es/ss/Satellite?L=es_ES&c=INESeccion_C&cid=1259925408327&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout&param1=PYSDetalle&param3=1259924822888

Figure 1: Distribution of maintenance payments (MP) and family home (FH) attribution/suppression requests and resolutions for each plaintiff gender



4 DIRECT GENDER EFFECT

To assess the direct influence of gender features on court decisions, we train a collection of classification models and use feature importance analysis. This analysis is motivated by the tradition in Explainable AI (XAI) to use a surrogate model to understand the decision-making process and determine the importance of features in relation to outcomes [24]. In our work, we use models to analyze historical decisions. We train a court decision prediction model that allows us to mimic the decision-making process of a judge. Furthermore, we train predictors for the gender of both the plaintiff and the judge to determine which features are most related to the gender variables.

This methodology may contain some degree of error as it is an approximation. Nonetheless, it is a well-established and widely used technique that enables us to quantify and simplify the reasoning process.

4.1 Feature importance analysis

Feature importance assigns a score to each feature based on their importance to predict the output. We study feature importance across three different ML classifiers. The features used by the classifiers are the features related to the court arguments listed in Table 3, the plaintiff's request features, and the gender variables. Each classifier aims to predict a different target variable:

(1) **Plaintiff gender**. Predicts the gender of the plaintiff involved in a sentence.
(2) **Judge gender**. Predicts the gender of the judge involved in a sentence.
(3) **Court decision**. Predicts the victory condition, specifically, whether the requested and resolved custody types are the same.

For training each ML classifier, we train five different models using cross-validation (5-fold) and select the one that yields better performance. The comparison is done between:

(1) **Logistic Regression**. random_state=0, C=1, max_iter=1000, solver='lbfgs'
(2) **Random Forest**. n_estimators = 1000
(3) **Support Vector Classification with linear kernel function**. gamma='auto'
(4) **Support Vector Classification with Radial Basis Function kernel**. gamma='auto'
(5) **AdaBoost classifier**. n_estimators=1000, random_state=0

Given that the dataset is not well balanced across genders or in terms of victory condition, we use balanced accuracy as the performance metric to evaluate model predictions. The balanced accuracy is computed as:

$$\frac{sensitivity + specificity}{2}$$

where the sensitivity is the true positive rate and the specificity is the true negative rate.

4.1.1 Plaintiff side. In the prediction of plaintiff gender, the Logistic Regression classifier performs the best, with a balanced accuracy of around 85%, indicating a significant difference in the characteristics of sentences involving male and female plaintiffs. However, when dividing the data by custody type request, the model's performance decrease, with a balanced accuracy of 58% for sole custody cases and 51% for joint custody cases. This suggests that the initial model is heavily relying on the request type feature, which is consistent with the strong correlation identified between plaintiff gender and request type in Section 3.3.

For the court decision prediction, the Random Forest classifier shows the best performance with a balanced accuracy of around 88%. Training the classifier without the plaintiff gender feature results in a similar accuracy, indicating that this feature is not crucial for the model's learning process. This conclusion is also reached when training models with data divided by request type.

For the same court decision classifier, we evaluate the disparities in feature importance coefficients across genders, with the data divided by request type. We create a disparity score to reveal the features with the greatest variation in coefficients for different genders, reflecting the features unequal impact on court decisions. We observe that for sole custody requests, only two arguments have a disparity higher than the threshold of 0.01, "Best interests of the child" and "Parents' readiness". In contrast, when looking at joint custody requests, seven arguments have a disparity higher than the threshold, indicating greater differences in the arguments used in joint custody cases compared to sole custody cases.

4.1.2 Judge side. With respect to the judge gender prediction, the classifier with the highest performance is the Ada Boost classifier with a balanced accuracy of 52%. This suggests that no classifier is able to effectively differentiate between judge genders. Similar performance was found when training the classifiers for the two types of custody.

In the analysis involving the court decision classifier, when the classifier is trained without the judge gender feature, the accuracy is similar, indicating that the judge gender is not highly important for the model.

Furthermore, when comparing the feature importance across genders, only "Circumstances of the children (fact)" has a disparity score above the threshold 0.01, being more important when the judge is female. This is also the only argument with disparity score

above threshold when performing the same evaluation with the data divided by request type. In these cases, the disparities are more pronounced, with a double disparity for sole custody cases and four times greater for joint custody cases.

5 INDIRECT GENDER EFFECT

In this section, we quantify the indirect effect of gender on court decisions using propensity score matching and exact matching. Although randomized controlled experiments are the best way to estimate causal effects of an intervention or treatment, in many cases - as ours -, we are under observational studies where the assignment of the intervention is not controlled or may not be randomized [21]. In such scenarios, propensity score matching and exact matching can be used to control confounding variables. We use both methods for the same evaluation to provide robustness to the results and conclusions [12]. Moreover, although propensity score matching is mathematically suitable, it is not fully suitable theoretically as gender is not a treatment we apply. Exact matching is more theoretically suitable for our case as it is more appropriate to evaluate the concept of inherent gender.

In our case, we consider the gender as the "intervention" and the winning label as the "effect". Males are considered the "treated group" and females the "control group" - as an analogy of usually privileged and unprivileged groups of the protected attribute. Our objective is to detect if there are confounding variables present i.e., if the arguments used by the court are related to both the winning probability and the probability of belonging to a particular gender [11]. If such results are found, it suggests that gender may be embedded in the court's arguments and influence the decision.

5.1 Propensity score matching

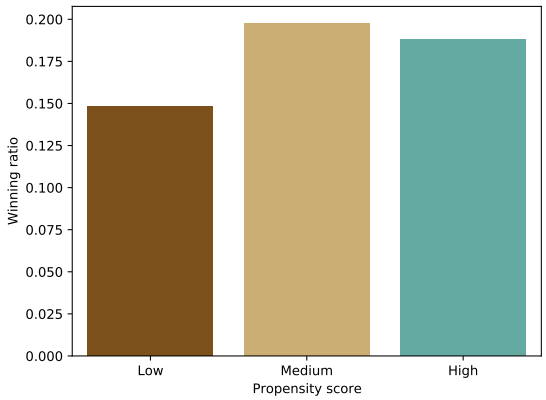
The propensity score is the probability of receiving a treatment given specific baseline characteristics and it acts as a balancing score to simulate a randomized controlled experiment [2].

In our propensity score analysis, we divide cases into buckets based on the probability of belonging to a gender group, specifically, the "treated" group consisting of males. This probability is calculated using the confidence of the gender classifiers discussed in Section 4.1. For each bucket, we determine the ratio of winning sentences to identify any patterns that may indicate an indirect relationship between the treatment (gender) and the outcome (winning label).

5.1.1 Plaintiff side. Figure 2 displays the results of the propensity score matching analysis for the plaintiff gender. As observed, no clear pattern is found. Although there is a rising trend appearing in the first two buckets, it does not persist for the third, which prevents us from drawing strong conclusions about the relationship between the plaintiff's gender and the likelihood of winning using the propensity score matching method. Other visualizations with different amount of buckets lead to similar conclusions.

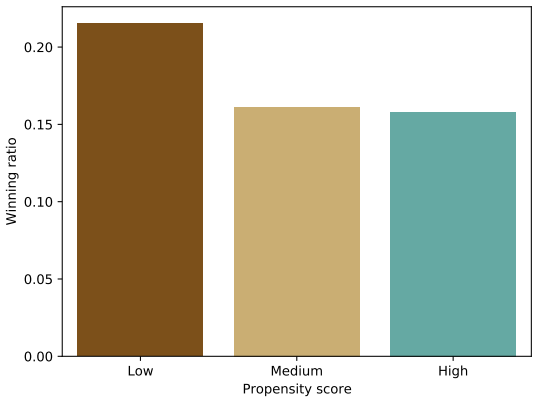
5.1.2 Judge side. Figure 3 shows a clear trend in the correlation between the judge's gender and the winning label. In particular, the buckets with cases that are more likely to be presided over by a female judge seem to have a higher ratio of winning sentences.

Figure 2: Propensity score plaintiff gender



Other visualizations with different amount of buckets lead to the same conclusions.

Figure 3: Propensity score judge gender



5.2 Exact matching

In exact matching, matched samples are created to intend to replicate a randomized experiment and balance the distribution of covariates i.e., variables that may affect the final outcome (excluding the actual treatment), between the treated and control groups [23].

The process we follow to create the matched samples is illustrated in Figure 4. We start by dividing the data into two groups using the judge ID, which is assumed to be random and allows us to avoid double matches. Group 0 includes sentences with even judge IDs and group 1 includes sentences with odd judge IDs. Then, we extract a sample A from group 0, consisting of sentences with female plaintiffs. Then, for each sentence $x \in A$, we search for a male plaintiff sentence y in group 1 that minimizes the distance $d(x, y)$. Finally, we select the 25% of most similar sentences to create the collection of matched samples.

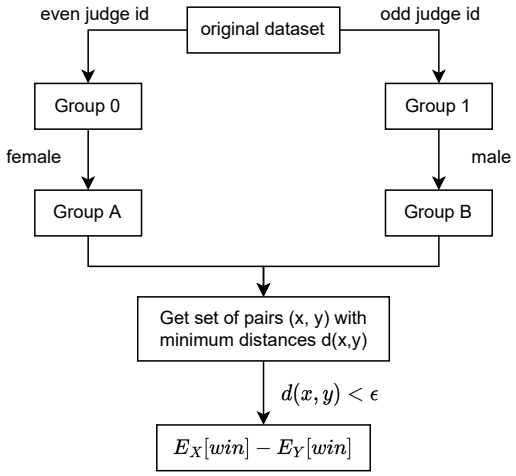
The distance $d(x, y)$ between each pair is computed using three different matching criteria:

(1) $\mathbf{D_1}$. Euclidean distance considering all the arguments used by the court as well as all the request variables.
(2) $\mathbf{D_2}$. Weighted euclidean distance with the most important variables in the probability of winning, where the weights are proportional to the importance for the probability of winning.
(3) $\mathbf{D_3}$. Weighted euclidean distance with only the 10 most important variables in the probability of winning, where the weights are proportional to the importance for the probability of winning.

The most relevant variables and their weights are extracted from the feature importance analysis. More precisely, from the coefficients in the court decision predictor. The judge id and the gender of the plaintiff is not considered in any of the distances. This analysis is done separately for each type of custody request: sole and joint.

The gender effect, once matched samples are extracted, is computed as $E_X[win] - E_Y[win]$, where X is the sample containing female plaintiffs and Y the sample containing male plaintiffs.

Figure 4: Exact matching methodology



5.2.1 Plaintiff side. The results for the exact matching technique for the plaintiff side are presented in Figure 5. We can see that for sole custody requests, the female sample has slightly higher victory rates, indicating that a female is more likely to win a sole custody case. The same results are found for all similarity functions. This effect is not observed (or appears reversed, with much higher differences) in joint custody cases. Different results are found for different similarity functions, making the insights less robust. We hypothesize that the observed differences in sole custody sentences are caused by the affirmative that females tend to win sole custody cases at a higher rate as seen in Section 3.3.1.

5.2.2 Judge side. Figure 6 displays the results of the exact matching technique applied to the judge side. As depicted, all similarity functions show positive effects for female judges, for both sole and joint request types. The strongest effects are observed for *D2*, which utilized the most important features for calculating similarity.

The results suggest that female judges are more likely to grant the requested custody type, resulting in higher victory rates when a female judge is assigned to a case.

Figure 5: Differences in the victory rates between matched groups (plaintiff)

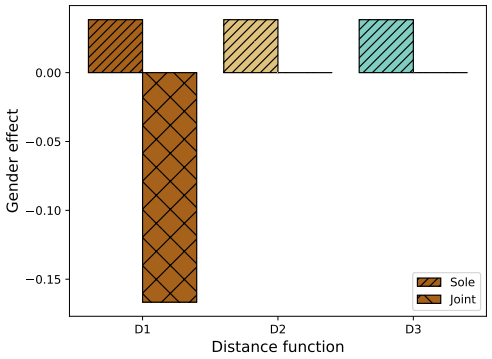
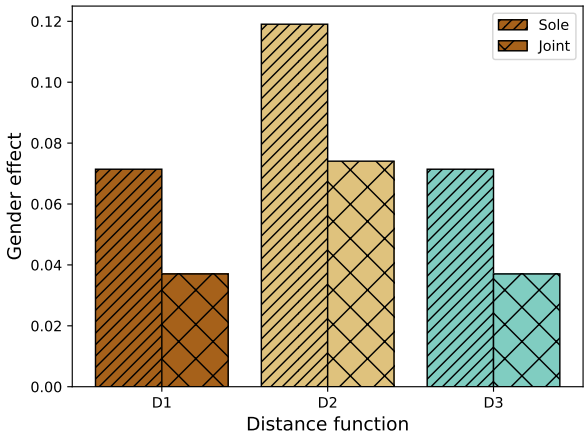


Figure 6: Differences in the victory rates between matched groups (judge)



6 DOCTRINE UNIFICATION

"*Recurso de casación para unificación de doctrina*" is a legal term in Spain meaning "appeal for the unification of doctrine", which can be used in child custody matters. It is an extraordinary appeal whose purpose is to promote consistency and coherence in the application of the law. It is used when legal interpretations are contradictory with previous judgments involving the same litigants or similar

parties, concerning the same circumstances and based on similar facts, arguments, and claims.[4]

In line with this concept, we conduct an analysis to identify cases where this appeal could be used and examine the influence of gender. Our hypothesis is that if there exist similar cases that only differ in gender features but have different resolutions, it could reveal evidence of bias.

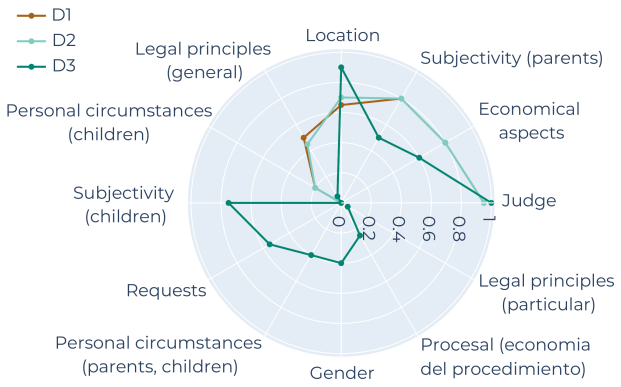
In order to find those similar cases with different resolutions we use counterfactual fairness. In the context of counterfactual fairness, a decision is considered fair for an individual when it remains unchanged for both the individual and their counterfactual counterpart, which is an alternate version of the same individual who belongs to a different demographic group [19].

To find "candidates for doctrine unification" we seek for the most similar sentences that differ in the court decision. We use various similarity functions for robustness purposes. The three distance functions employed are the same to those previously discussed in detail in Section 5.2. To select the candidates, we sort the pairs of sentences by their similarity according to each similarity function and then use an ad-hoc threshold to select a reasonable number of sentences that could be manually evaluated (20 sentences for each similarity function). Finally, we analyze the frequency of the differing features between the selected subset of most similar sentences with different label.

6.1 Doctrine unification results

Figure 7 displays the frequency at which feature categories differ among the 20 most similar sentences for each of the three aforementioned distance computation methods. The plot corresponds to a radar chart where each label represents a category. Values indicate the portion of sentences differing in each feature category, for each of the similarity functions analyzed. The categorization of court arguments is detailed in Table 4.

Figure 7: Differing feature categories



[4]"Recurso de casación para unificación de doctrina"(appeal for unification of doctrine): https://guiasjuridicas.laleynext.es/Content/Documento.aspx?params=H4sIAAAAAAAEAMtMSbF1jTAAAkNDC1MjU7Wy1KLizPw8WyMDAwsDcwMLkEBmWqVLfnJIZUGqbVpiTnEqAN4cKl41AAAAWKE

In our results, the gender factor only emerges as a difference when using the distance function $D3$. Further investigation reveals that this gender difference corresponds to the judge's gender, which is found to differ in 8 out of the 20 most similar sentences computed with $D3$. However, the gender factor does not appear as a difference when using $D1$ and $D2$. Other categories that vary consistently across all three distance functions include "Economic aspects", "Subjectivity (parents)", "Legal principles (general)", "Location" and the judge's ID.

It's worth noting that while the last two features may suggest doctrine variations across locations and among judges, we cannot confirm that these variations are solely caused by these features. All the identified pairs of most similar cases also differ in at least one other category of legal principles and/or facts, which could be a valid reason for the judge to assign different outcomes.

7 CONCLUSIONS AND FUTURE WORK

In our study, we investigate gender biases in the context of child custody sentences in Spain. To do this, we analyze a dataset containing manually-labeled information from over 1,800 cases. Our findings reveal significant differences in economic factors, with female plaintiffs more frequently requesting the attribution of maintenance payments and the family home, while male plaintiffs are more likely to request their suppression (see Table 1 and Table 2).

Although the direct influence analysis did not reveal a strong correlation between the gender of the plaintiff or the judge and the final outcome, it does indicate significant differences in the features present in sentences for male plaintiffs compared to female plaintiffs.

Our results from the indirect effect analysis do not reveal a strong or systematic difference between genders of the plaintiffs. However, our findings reveal a significant correlation between the gender of the judge and the outcome of the child custody cases in Spain. This is a noteworthy result that merits further investigation and attention, as it highlights the potential for gender biases in the judicial system. It is crucial to understand the underlying causes of these disparities to ensure fairness and equity in the legal process. To address this issue, it may be beneficial to implement measures such as bias detection, mitigation and training to prevent any potential biases in the future.

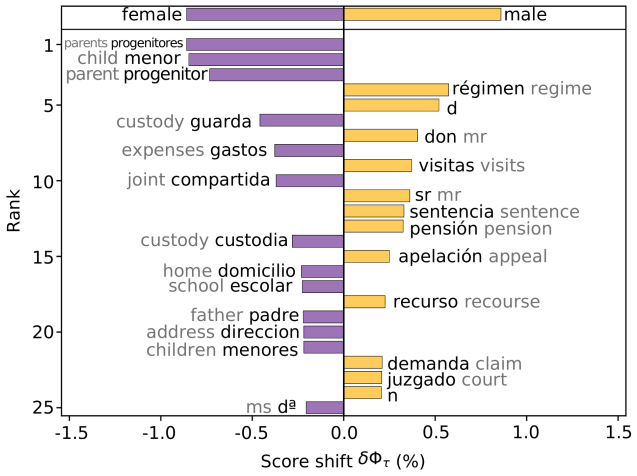
7.1 Policy implications

The obtained results leave an open door for an abuse, where trying to be assigned with a female judge could increase the winning chances for a given plaintiff in most of the cases. This section discusses potential policy implications of the observed disparities.

7.1.1 Judge assignation. Although we know that judges are assigned randomly, the size and consistency of observed disparities suggest that it is insufficient. The fact that even with a random assignation the observed disparities are present, makes them even more significant, suggesting the need for other ways of mitigating their potential effects.

7.1.2 Need for more unified criteria. An analysis of how the usage of certain terminology varies among sentences corresponding to judges of different genders suggest different criteria between them.

Figure 8: Terms proportion shift between judge gender sentences



To evaluate the differences in the doctrine among judge genders we use Proportion Shift. This technique highlights those terms whose relative frequency varies the most between two corpus of text. With $p_i^{(1)}$ being the frequency of word i in the first corpus and $p_i^{(2)}$ in the second corpus, the Score shift is calculated as $\delta p_i = p_i^{(2)} - p_i^{(1)}$. Higher values of δp_i indicate terms whose frequency varies the most between judge gender.

In our analysis, the first corpus contains the texts from female judges and the later contains the text of male judge sentences.

As can be seen in the Proportion Shift plot depicted in Figure 8, female judges use terminology related to family structure with higher rates (using terms such as “parents”, “child”, “custody”, “family home”, among others). In the case of male judges, their sentences seem to reflect higher rates of legal anchors usage (with terms such as “regime”, “visits”, “recourse”, among others).

This suggests that there is still space for the implementation of stronger doctrine unification approaches that could help to ensure an equal doctrine among judges of different gender.

7.2 Limitations and future work

This research utilizes a substantial dataset of samples, which is a noteworthy accomplishment given the complexity of annotating this type of data. However, it is important to note that this dataset is only a subset of all the sentences issued in Spain and there is no information on the sampling methodology used. This raises the possibility that unintended biases may be present in the results derived from the data.

Furthermore, the dataset is comprised of second-instance sentences from the specific domain of child custody in Spain, which may limit the generalizability of the findings to first-instance sentencing, other domains, geographic regions, and/or cultures. Additionally, more research could be done on the effects of other demographic factors, in case they were made available. Other demographic factors that could be studied include the age, race, education level, and socioeconomic status of both the plaintiffs and judges. Additionally, studying the impact of the lawyer’s demographic characteristics on the outcome of the trial could also be an interesting area of research. Another factor that could be studied is the length of service of the judge and how it affects the decision making. Furthermore, it would be interesting to explore how the gender composition of the court panel affects the outcome of the trial and whether there are any significant differences between solo, two-judge and three-judge panels.

7.3 Reproducibility

Open-source code and the data generated and used throughout the research are available at: https://github.com/juliariera/gender-disparities-in-child-custody-sentencing-in-Spain

8 ACKNOWLEDGEMENTS

We appreciate the effort taken by researchers of the Bidaraciv project on the collection of the analyzed dataset.

D. Solans acknowledges that this work has been partly funded by European Union Horizon 2020 program under grant agreement No. 101021808 (SPATIAL-H2020 project).

C. Castillo acknowledges that this work has been partially supported by: "la Caixa" Foundation (ID 100010434), under agreement LCF/PR/PR16/ 51110009; EU-funded projects "SoBigData++" (grant agreement 871042) and "FINDHR" (grant agreement 101070212).

REFERENCES

[1] Kevin Ashley. 2019. A brief history of the changing roles of case prediction in ai and law. *Law in Context. A Socio-legal Journal*, 36, (Aug. 2019), 93–112. DOI: 10.26826/law-in-context.v36i1.88.

[2] Peter C. Austin. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46, 3, (June 2011), 399–424. DOI: https://doi.org/10.1080/00273171.2011.568786.

[3] Leena Autonen-Vaaraniemi. 2022. Family professionals’attitudes and stance-taking on post-divorce fatherhood: a qualitative attitude approach. *International Journal of Child, Youth and Family Studies*, 13, 1, 56–81.

[4] Floris Bex and Henry Prakken. 2021. On the relevance of algorithmic decision predictors for judicial decision making. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, 175–179.

[5] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (Proceedings of Machine Learning Research). Sorelle A. Friedler and Christo Wilson, (Eds.) Vol. 81. PMLR, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html.

[6] Xiqian Cai, Pei Li, Yi Lu, and Hong Song. 2021. Gender in-group bias: evidence from judicial decisions. *Social Science Research Network*, (July 2021). DOI: http://dx.doi.org/10.2139/ssrn.3869294.

[7] Luiza Lopes Franco Costa, Ana Beatriz Dillon Esteves, Roxana Kreimer, Noel Struchiner, and Ivar Hannikainen. 2019. Gender stereotypes underlie child custody decisions. *European Journal of Social Psychology*, 49, 3, 548–559. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2523. DOI: https://doi.org/10.1002/ejsp.2523.

[8] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108, 17, 6889–6892. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1018033108. DOI: 10.1073/pnas.1018033108.

[9] Laboratorio Jurídico-Empresarial de la Universidad de Zaragoza. 2021. Bidaraciv. (Sept. 2021). Retrieved September 30, 2022 from https://github.com/labje/bidaraciv.

[10] ITAINNOVA Instituto Tecnológico de Aragón. 2020. Webinar "la inteligencia artificial y la actividad judicial. (Sept. 2020). Retrieved September 30, 2022 from https://www.youtube.com/watch?v=zrXOLBzetys.
[11] Science Direct. [n. d.] Confounding variable. (). Retrieved September 30, 2022 from https://www.sciencedirect.com/topics/nursing-and-health-professions/confounding-variable.
[12] Anne Burden et al. 2017. An evaluation of exact matching and propensity score methods as applied in a comparative effectiveness study of inhaled corticosteroids in asthma. *Pragmatic and observational research*, 2017, 8, (Mar. 2017), 15–30. DOI: https://doi.org/10.2147/POR.S122563.
[13] Moritz Hardt. 2016. How big data is unfair. (July 2016). https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de.
[14] Rut Iturbide, Patricia Amigot, and Susana Covas. 2021. Experiencias de mujeres en procesos de separación y divorcio: un estudio cualitativo sobre dinámicas de poder masculino y violencias naturalizadas. *Instituto Navarro para la Igualdad/Nafarroako Berdintasunerako Institutua*. https://www.igualdadnavarra.es/es/estudio-experiencias-de-mujeres-en-procesos-de-separacion-y-divorcio.
[15] Lauren Kirchner Jeff Larson Surya Mattu and Julia Angwin. 2016. How we analyzed the compas recidivism algorithm. (Sept. 2016). Retrieved September 30, 2022 from https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.
[16] Jerry Kang, Mark Bennett, Devon Carbado, Pam Casey, and Justin Levinson. 2011. Implicit bias in the courtroom. *UCLa L. rev.*, 59, 1124.
[17] Marzieh Karimi-Haghighi and Carlos Castillo. 2021. Enhancing a recidivism prediction tool with machine learning: effectiveness and algorithmic fairness. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (ICAIL '21). Association for Computing Machinery, São Paulo, Brazil, 210–214. ISBN: 9781450385268. DOI: 10.1145/3462757.3466150.
[18] Kathryn M Kroeper, Victor D Quintanilla, Michael B Frisby, Nedim Yel, Amy Applegate, Mary C Murphy, and Steven Sherman. 2020. (Feb. 2020). DOI: 10.1037/law0000229.
[19] Chris Russell Matt J. Kusner Joshua Loftus and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, (Eds.) Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.
[20] Ines P Murillo Huertas, Raúl Ramos, and Hipólito Simon. 2017. Regional differences in the gender wage gap in spain. *Social Indicators Research*, 134, 3, 981–1008.
[21] Rachel B. Jimenez Nafisha Lalani and Beow Yeap. 2020. Understanding propensity score analyses. *International journal of radiation oncology, biology, physics*, 107, 3, (July 2020), 404–407. DOI: https://doi.org/10.1016/j.ijrobp.2020.02.638.
[22] José Félix Muñoz Soro and Carlos Serrano-Cinca. 2021. A model for predicting court decisions on child custody. *PloS one*, 16, 10, (Oct. 2021), 21. DOI: https://doi.org/10.1371/journal.pone.0258993.
[23] Elizabeth A. Stuart. 2010. Matching methods for causal inference: a review and a look forward. *Statistical science*, 25, 1, (Feb. 2010), 1–21. DOI: https://doi.org/10.1214/09-STS313.
[24] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: a survey on methods and metrics. *Electronics*, 10, 5, 593. DOI: https://doi.org/10.3390/electronics10050593.

Table 3: List of court arguments divided by legal principles and facts: code and description.

Legal principles	
LP_BI	Best interests of the child
LP_PE	Parents equality
LP_RA	Proportionality in the responsibilities
LP_RJ	Res judicata
Facts	
FT_RP	Psycho-social report
FT_CW	Opinion of children
FT_CC	Circumstances of children
FT_CR	Roots of children
FT_RA	Parents' relationship and attitude
FT_RD	Parents' readiness
FT_PD	Parents' previous dedication
FT_AG	Parents' agreements

Table 4: Categorization of court arguments.

Feature	Category
LP_BI	Legal principles (general)
LP_PE	Legal principles (general)
LP_RA	Economical aspects
LP_RJ	Procesal (economia del procedimiento)
FT_RP	Personal circumstances (parents, children)
FT_CW	Subjectivity (children)
FT_CC	Personal circumstances (children)
FT_CR	Subjectivity (children)
FT_RA	Subjectivity (parents)
FT_RD	Economical aspects
FT_PD	Subjectivity (parents)
FT_AG	Subjectivity (parents)

Received March 2023; accepted 11 April 2023