Algorithmic Fairness and Efficiency in Targeting Social Welfare Programs at Scale

Alejandro Noriega†
MIT Media Lab
Cambridge, MA

Bernardo Garcia-Bulle†
ITAM
Mexico City, Mexico

Luis Tejerina
IADB
Washington, DC

Alex Pentland
MIT Media Lab
Cambridge, MA

ABSTRACT

Targeted social programs, such as conditional cash transfers (CCTs), are a major vehicle for poverty alleviation throughout the developing world. Only in Mexico and Brazil, these reach nearly 80 million people (25% of population), distributing +8 billion USD yearly. We study the potential efficiency and fairness gains of targeting CCTs by means of artificial intelligence algorithms. In particular, we analyze the targeting decision rules and underlying poverty prediction models used by national-wide CCTs in three middle-income countries (Mexico, Ecuador, and Costa Rica). Our contribution is three-fold: 1) We show that, absent explicit measures aimed at limiting algorithmic bias, targeting rules can systematically disadvantage population subgroups, such as incurring exclusion errors 2.3 times higher on poor urban households compared to their rural counterparts, or exclusion errors 2.2 times higher on poor elderly households compared with poor traditional nuclear families. 2) We constrain the targeting algorithms towards achieving fairness, and show that, for example, mitigating urban/rural unfairness in Ecuador can imply substantial costs in overall accuracy, yet, we also show that in the case of Mexico mitigating unfairness across four different types of family structures can be achieved at no significant accuracy costs. 3) Finally, we provide an interactive decision-support platform that allows even non-expert stakeholders to explore the space of possible AI-based decision rules, visualize their implications in terms of efficiency, fairness, and their trade-offs; and ultimately choose designs that best fit their preferences and context.

Keywords: Algorithmic fairness; Targeting social programs; Prediction for policy; Trade-off space exploration

† Authors contributed equally to this work.
* This paper is work in progress. Content additions and improvements in synthesis, analysis and writing are yet to be implemented throughout.

Bloomberg Data for Good Exchange Conference.
Sep-2018, New York City, NY, USA.

1. INTRODUCTION

As automated decision-making systems have become increasingly ubiquitous—e.g., in criminal justice[18], medical diagnosis and treatment[17], human resource management[7], social work[12], credit[15], and insurance—there is widespread concern about how these can deepen social inequalities and systematize discrimination [21, 20]. Consequently, substantial work on defining, measuring, and optimizing for algorithmic fairness has surged in recent years. This rising field of research has focused on offline domains such as the criminal justice system [9, 4], child maltreatment hotlines[10], and predictive policing[22]; as well as online domains such as targeted advertising[23], search engines[13], and face recognition algorithms[5].

Targeted social welfare programs. The present work focuses on targeted social welfare programs, which encompass some of today's largest algorithmic decision-making systems in offline domains, and whose decisions bare substantial impact on the the lives of millions of people worldwide. In particular, we focus on conditional cash transfer programs (CCTs), which provide a financial stipend to families in poverty, and require them to comply with "co-responsibilities", such as maintaining children in school, and attending regular medical appointments[11].

CCTs are a major vehicle for poverty alleviation in the developing world. There are more than 100 national CCTs worldwide (Figure S1 for world map of national CTTs)[2]. Only in Mexico and Brazil, for example, these reach nearly 80 million people (25% of the population), distributing +8 billion USD yearly (0.3% of GDP)[16].

CCTs are targeted in the sense that only a subset of the population, generally those below a specified poverty line, is eligible to programs' benefits. However, reliable income data is typically not available and costly to procure, as households

	Poverty ratio	N	No. of Variables
Mexico	35%	70,311	183
Ecuador	25%	30,338	126
Costa Rica	21%	10,711	165

Table 1: **Household survey data statistics.** This table presents the poverty ratio in the households in the survey, and the number of observations in each survey.

in the target population participate mainly in informal economic sectors. Hence targeting of CCTs most commonly relies on poverty prediction algorithms that decide households' eligibility based on observable and less costly proxy data, such as education levels, demographics, and the assets and services of households[16].

Fairness and Efficiency in CCTs. Substantial previous literature has looked into the accuracy of different targeting methods[1, 8]. Yet, the potential existence of algorithmic unfairness—in terms of how these inference systems might differentially affect population subgroups—and the ways in which program managers can mitigate disparities, have not been thoroughly studied.

Summary of contributions. The present work shows quantitatively how 1) substantial disparities across subgroups may be introduced by the targeting systems of CCTs, 2) disparities can be effectively mitigated by constraining algorithms towards fairness, and 3) that such constraints imply costs in terms of the overall targeting accuracy of the system (inclusion and exclusion errors). Finally, we propose an AI-based decision-support tool for helping program managers navigate the space of possible decision rules, in terms of their performance and trade-offs across accuracy and fairness.

2. DATA

For this study, we use data from three countries to train and test our models. The countries are Costa Rica, Ecuador and Mexico. We chose these countries as they represent large (121M), medium (16M) and small (4M) populations, and that is reflected in the number of observations in our datasets. Despite the differences, our methods worked in all of them.

Household surveys data. We used household surveys data as those are the publicly available best proxies for household's income. Household surveys are applied periodically in all our three countries to a representative sample of the population.

These surveys are extensive, which translates into reliable detailed information. As an example, Mexico's council on evaluation of social development policies (CONEVAL) uses ENIGH, the same survey we use, to measure poverty in the country. Table 1 presents basic details on the surveys.

The surveys include the total income per household, demographic variables, and variables related to the physical state of the household and the objects it contains. Examples of the former are education levels, ages, and genders. Examples of the latter are construction materials, number of rooms, utilities available, cars, and appliances.

3. POVERTY PREDICTIONS

Prediction algorithms

In order to estimate poverty we implement a series of prediction algorithms. The algorithms take as input the series of variables described in Section 2, and use those to estimate whether households are below or above the poverty line. We trained the algorithms with 75% of the data, and measured all performance outcomes with the remaining 25% (test set).

While testing algorithms from ML and traditional regressions, gradient boosted trees yielded the best results in test sets across all countries. Hence all results in what follows of this paper were computed based on it. A full comparison of performance across algorithms is beyond the scope of this paper.

The output of classification algorithms are probability scores on the likelihood of a household being poor. Targeting decisions are based on those scores. A score threshold is set by program managers, above which households are consider eligible for the program.

Results

Imperfections in poverty estimations will lead to two kinds of errors: exclusion and inclusion errors. The first refers to the proportion of actually poor people who did not receive the aid whilst the second refers to the proportion of aided households which were in fact above the poverty line, therefore not eligible. We use these errors to assess quality in the estimator. There is a natural trade off between these two. If a program wants to lower its exclusion error, by increasing the number of receivers, it will tend to take households which are increasingly likely to be non-poor (as the likeliest are the first to be selected). This will result in higher inclusion errors. In the extremes, having zero beneficiaries means 0% inclusion error and 100% exclusion error; and having the whole population receive the program will translate into the maximum inclusion error (equal to the proportion of non-poor people in the population) and 0% exclusion error.

Once a threshold over poorness scores is decided, households above the threshold will receive the program and houses below will not. Increasing the threshold will lower the inclusion error while also increasing the exclusion error, forming the exclusion-inclusion trade-off. Figure1 shows results of computing such trade-off for the three countries in this study.

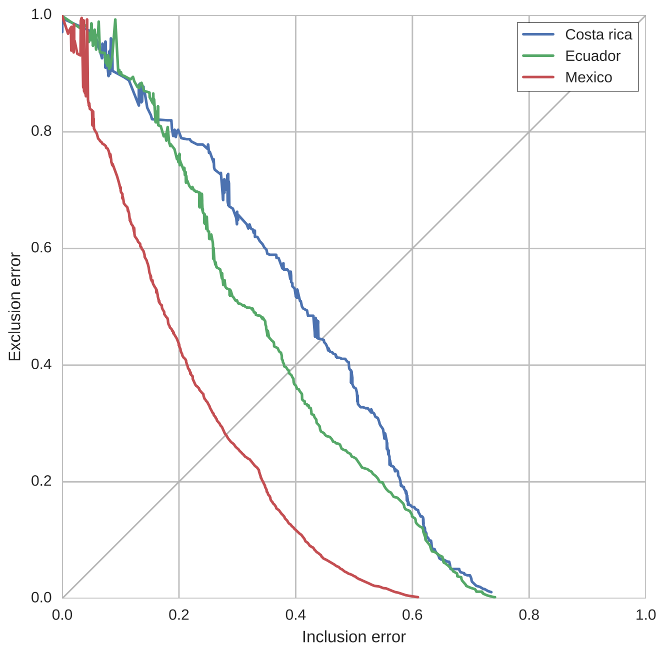


Figure 1: Inclusion and exclusion errors.

A line is plotted for each country, showing the exclusion error entailed by each achievable level of inclusion error. For example, we see that Mexico's CTT can achieve an exclusion error as low as 10% while incurring an inclusion error of 40%.

4. DISPARATE IMPACT AND THE COST OF FAIRNESS

Notions of fairness. Several notions of fairness and their corresponding formalizations have been proposed, most of which require that statistical properties hold across two or more population subgroups. *Demographic* or *statistical parity* requires that decisions are independent from group membership [6, 24, 19], such that $P\{\hat{Y} = 1|A = 0\} = P\{\hat{Y} = 1|A = 1\}$, for the case of binary classification and sensitive attribute $A \in \{0, 1\}$. Most recent work focuses on meritocratic notions of fairness, or *error rate matching* [14, 3], such as requiring population subgroups to have equal false positive rates (FPR), equal false negative rates (FNR), or both (*equalized odds*), i.e., $P\{\hat{Y} = 1|A = 0, Y = y\} = P\{\hat{Y} = 1|A = 1, Y = y\}$, $y \in 0, 1$. In this work we adopt the latter set of fairness notions. Refer to [25] for a survey on computational measures of fairness.

In the context of CCTs, we measure fairness of targeting systems in terms of how similar their exclusion errors are across population subgroups. As CCTs provide a positive benefit, this measure of fairness captures the notion of *equality of opportunity*[14], where, for example, we require poor single-parent households to have the same probability of accessing the social program as poor traditional nuclear families.

Observed unfairness. Figure 2 shows the distribution of exclusion errors across population subgroups—urban/rural, and family type subgroups—when applying the poverty prediction algorithm without fairness constraints, and compares it against its fairness-constrained modification. It is shown that unconstrained algorithms can incur substantially different exclusion errors to different population subgroups. For example, poor urban households in Ecuador are 2.3 times more likely to be excluded from the program than their rural counterparts. Similarly, poor elderly households in Mexico are 2.2 times more likely to be excluded than their traditional nuclear family counterparts.

Mitigating unfairness. We introduced fairness constraints by establishing differentiated thresholds for each population subgroup, and adjusting them to guarantee parity in exclusion errors across subgroups, i.e., guaranteeing equality of opportunity[14]. The second column of Figure2 shows the results (note that all methods are calibrated on the training set, and results computed on the test set). We see that the method was effective in mitigating exclusion disparities across subgroups. For example, exclusion error rates of poor urban and rural households in Ecuador were equalized, as well as exclusion errors across family types in Mexico.

The cost of fairness.

We measure the cost of fairness as the performance degradation (which can be translated either as more exclusion error, more inclusion error, or both) resulting from constraining a targeting system to achieve similar exclusion errors across population subgroups. Figure 3 shows the cost of a chieving fairness for the three countries analyzed. It shows that the cost varies both across the protected attribute and country context: for some cases the cost is negligible, such as between rural and urban groups in Mexico and Costa Rica; yet, in other cases costs can be considerable, such as between rural and urban groups in Ecuador (increasing both errors from 38% to 44%). Lastly, we see encouraging results in cases like unfairness between family types in Mexico and Costa Rica, where rather stark disparities can be mitigated with moderate costs (increasing both errors from 28% to 30% in Mexico, and increasing both errors from 44% to 46% in Costa Rica).

5. CONCLUSIONS AND FUTURE WORK

The present work shows that, absent explicit measures for mitigating unfairness, targeting algorithms of social welfare programs can introduce relevant and systematic disparities in the exclusion errors of population subgroups (i.e., urban/rural and family structure types). We also show that disparities can be attenuated by introducing fairness constraints, yet these may come at significant costs in overall prediction efficiency across the population.

This work highlights the relevance of making transparent for key decision-makers both the accuracy and fairness implications of the space of possible algorithmic designs, as well as the trade-offs among these. We developed an AI-based decision-support tool to aid program managers in navigating such space of possibilities, and making informed choices. We look forward to demo the tool at D4GX 2018.

6. REFERENCES

[1] V. Alatas, A. Banerjee, R. Hanna, B. A. Olken, and J. Tobias. Targeting the poor: evidence from a field experiment in indonesia. *American Economic Review*, 102(4):1206–40, 2012.

[2] W. Bank. World databank. *World development indicators*, 2015.

[3] Y. Bechavod and K. Ligett. Learning fair classifiers: A regularization-inspired approach. *arXiv preprint arXiv:1707.00044*, 2017.

[4] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: the state of the art. *arXiv preprint arXiv:1703.09207*, 2017.

[5] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[6] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *Data mining workshops, 2009. ICDMW'09. IEEE international conference on*, pages 13–18. IEEE, 2009.

[7] A. Chalfin, O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, and S. Mullainathan. Productivity and selection of human capital with machine learning. *American Economic Review*, 106(5):124–27, 2016.

[8] D. Coady, M. Grosh, and J. Hoddinott. The targeting of transfers in developing countries: review of experience and lessons. *Social Safety Net Primer Series, World Bank, Washington DC*, 2002.

[9] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[10] S. Cuccaro-Alamin, R. Foust, R. Vaithianathan, and E. Putnam-Hornstein. Risk assessment and decision making in child protective services: Predictive risk modeling in context. *Children and Youth Services Review*, 79:291–298, 2017.

[11] A. De Brauw and J. Hoddinott. Must conditional cash transfer programs be conditioned to be effective? the impact of conditioning transfers on school enrollment in mexico. *Journal of development Economics*, 96(2):359–370, 2011.

[12] P. Gillingham. Predictive risk modelling to prevent child maltreatment and other adverse outcomes for service users: Inside the âĂŸblack boxâĂŹof machine learning. *The British Journal of Social Work*, 46(4):1044–1058, 2015.

[13] E. Goldman. Revisiting search engine bias. *Wm. Mitchell L. Rev.*, 38:96, 2011.

[14] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016.

[15] C.-L. Huang, M.-C. Chen, and C.-J. Wang. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4):847–856, 2007.

[16] P. Ibarrarán, N. Medellín, F. Regalia, M. Stampini, S. Parodi, L. Tejerina, P. Cueva, and M. Vásquez. How conditional cash transfers work. 2017.

[17] J. Kleinberg, J. Ludwig, S. Mullainathan, and Z. Obermeyer. Prediction policy problems. *American Economic Review*, 105(5):491–95, 2015.

[18] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[19] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

[20] S. U. Noble. *Algorithms of Oppression: How search engines reinforce racism.* NYU Press, 2018.

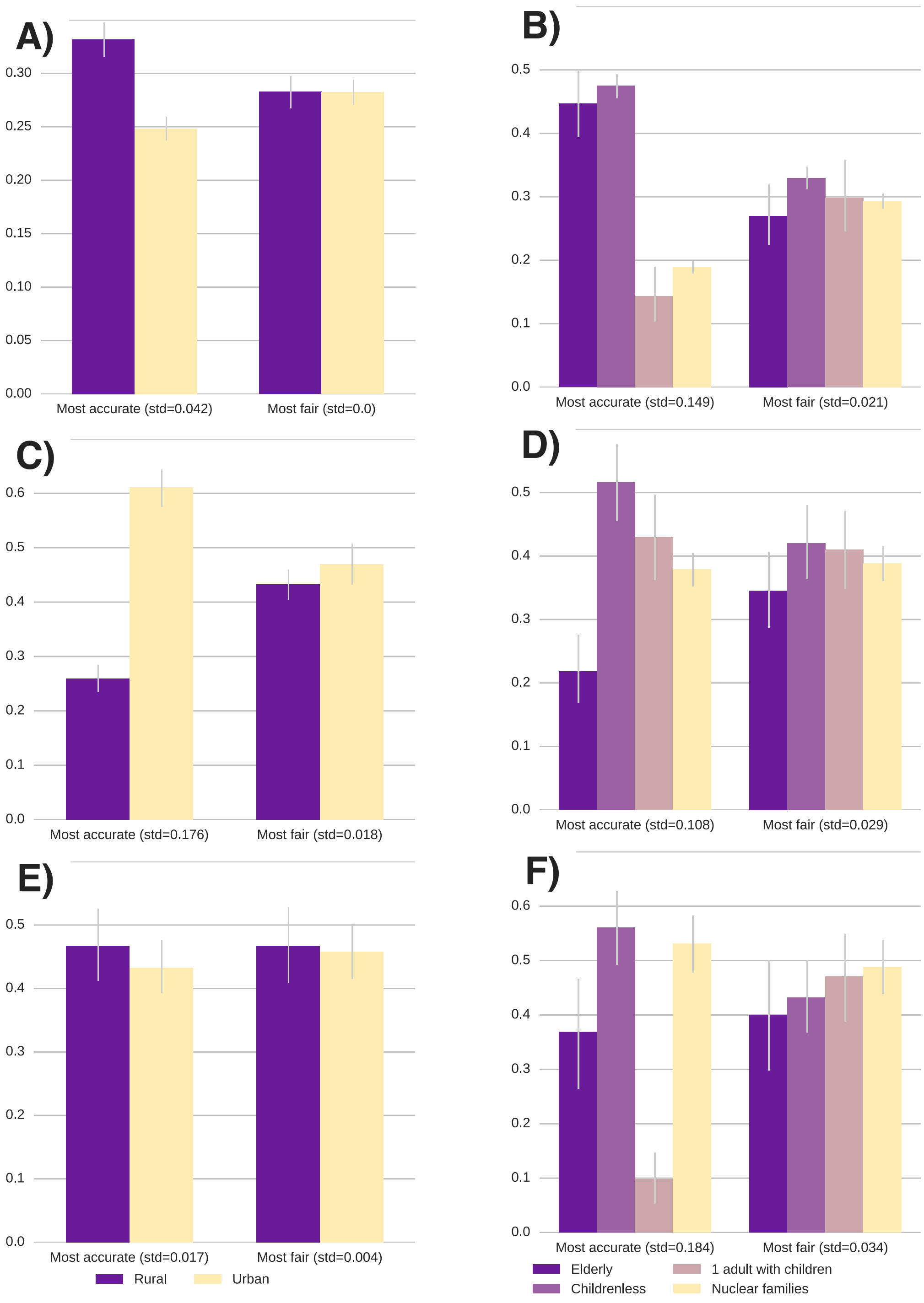
[21] C. O'Neil. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books, 2016.

[22] A. D. Selbst. Disparate impact in big data policing. *Ga. L. Rev.*, 52:109, 2017.

[23] L. Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.

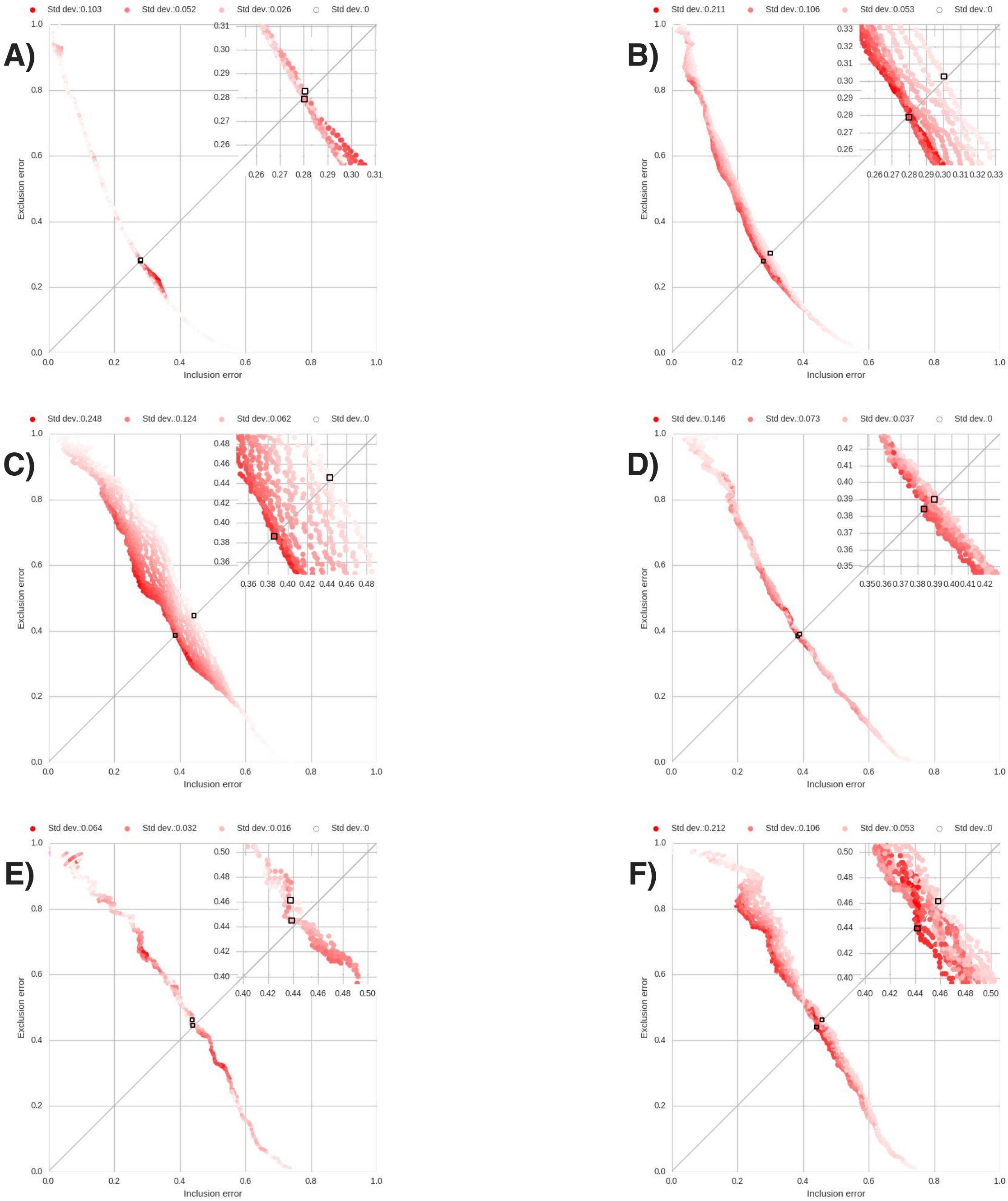
[24] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Learning fair classifiers. *arXiv preprint arXiv:1507.05259*, 2015.

[25] I. Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery*, 31(4):1060–1089, 2017.



Each group of bars relates to a thresholding strategy where we equal exclusion and inclusion error. The first row (A and B) has Mexico's data, the second (C and D) Ecuador's, and the third (E and F) Costa Rica's. The first column (A, C and E) measures inclusion in rural and urban groups, whilst the second (B, D and F) measures it for different types of families. On each plot we have two groups, one group is the most accurate thresholding strategy, and the other is the most fair thresholding strategy (both subject to the arbitrary restriction that inclusion and exclusion error must be equal). Both strategies set a threshold to each of the groups, however the first prioritizes accuracy and the second is constrained to having the same exclusion error per group. Although this is not exactly the case, this is due to the fact that the strategies were calculated on the training set, but the performance is shown on the test set. This shows how well these algorithms would generalize to new observations.

Figure 2: Inequality across population subgroups.



Each plot shows the inclusion, exclusion and unfairness (measured as standard deviation of the exclusion error amongst groups) and shown by coloring of different thresholding strategies. The first row (A and B) has Mexico's data, the second (C and D) Ecuador's, and the third (E and F) Costa Rica's. The first column (A, C and E) measures inclusion in rural and urban groups, whilst the second (B, D and F) measures it for different types of families. Each point in the plot is a thresholding strategy which can be either optimal, most fair, or something in between.

Figure 3: Unfairness, exclusion and inclusion errors, for thresholding strategies

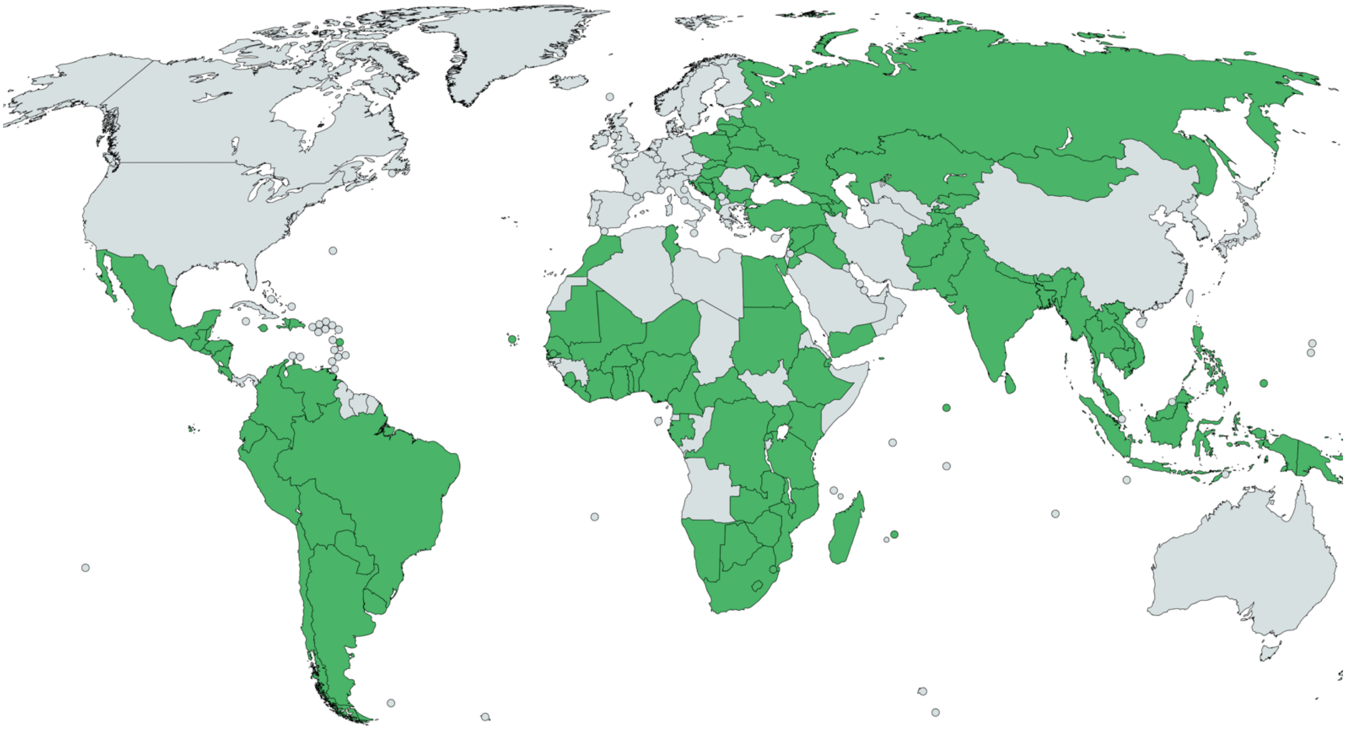


Figure 4: National cash transfer programs worldwide[2].