Demographic Bias in Machine Learning: Measuring Transference from Dataset Bias to Model Predictions

Iris Dominguez-Catena

October 2024

Department of Statistics, Computer Science and Mathematics,
Public University of Navarre (UPNA)
Supervisors: Mikel Galar Idoate, Daniel Paternain Dallo

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

*iris.dominguez@unavarra.es

Outline

Introducción

Objetivos

Propuestas

Conclusiones y trabajo futuro

Outline

Introducción

Objetivos

Propuestas

Conclusiones y trabajo futuro

Figure 1: A sample of FER2013/FER+, a popular FER dataset[1].

[1] Emad Barsoum et al. "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution". In: *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. Tokyo Japan, 2016, pp. 279–283.

Modalidades y aplicaciones de FER

Modalidades

- **Imagen** o video
- **RGB**, IR, Profundidad...
- Etiquetado **discreto** (emociones básicas de Ekman [2]) o continuo (NRC-VAD)...

Aplicaciones

- Multimedia interactivo
  - Películas emocionales
- Salud [3]
- Robótica asistiva [4]
- Seguridad pública [5]

---

[3] Paul Ekman and Wallace V. Friesen. "Constants across Cultures in the Face and Emotion.". In: *Journal of Personality and Social Psychology* 17.2 (1971), pp. 124–129

[4] Philipp Werner et al. "Automatic Recognition Methods Supporting Pain Assessment: A Survey". In: *IEEE Transactions on Affective Computing* 13.1 (2022), pp. 530–552

[5] Ritvik Nimmagadda, Kritika Arora, and Miguel Vargas Martin. "Emotion Recognition Models for Companion Robots". In: *J Supercomput* 78 (2022), pp. 13710–13727

[6] Luntian Mou et al. "Isotropic Self-Supervised Learning for Driver Drowsiness Detection With Attention-Based Multimodal Fusion". In: *IEEE Transactions on Multimedia* 25 (2023), pp. 529–542

Selección de datasets

3 **fuentes de datos**

- Lab: De laboratorio.
- ITW-I: De Internet.
- ITW-M: De películas y series.

Table 1: 20 datasets seleccionados.

Nombre	Año	Tipo	Imágenes	Vídeos	Sujetos
JAFFE	1998	LAB	213	—	10
KDEF	1998	LAB	4,900	—	70
CK	2000	LAB	8,795	486	97
Oulu-CASIA	2008	LAB	66,000	480	80
CK+	2010	LAB	10,727	593	123
GEMEP	2010	LAB	2,817	1,260	10
MUG	2010	LAB	70,654	—	52
SFEW	2011	ITW-M	1,766	—	330
FER2013	2013	ITW	32,298	—	—
WSEFEP	2014	LAB	210	—	30
ADFES	2016	LAB	—	648	22
FERPlus	2016	ITW	32,298	—	—
AffectNet	2017	ITW	291,652	—	—
ExpW	2017	ITW	91,793	—	—
RAF-DB	2017	ITW	29,672	—	—
CAER-S	2019	ITW-M	70,000	—	—
LIRIS-CSE	2019	LAB	26,000	208	12
iSAFE	2020	LAB	—	395	44
MMAFEDB	2020	ITW	128,000	—	—
NHFIER	2020	ITW	5,558	—	—

Fuentes de sesgo

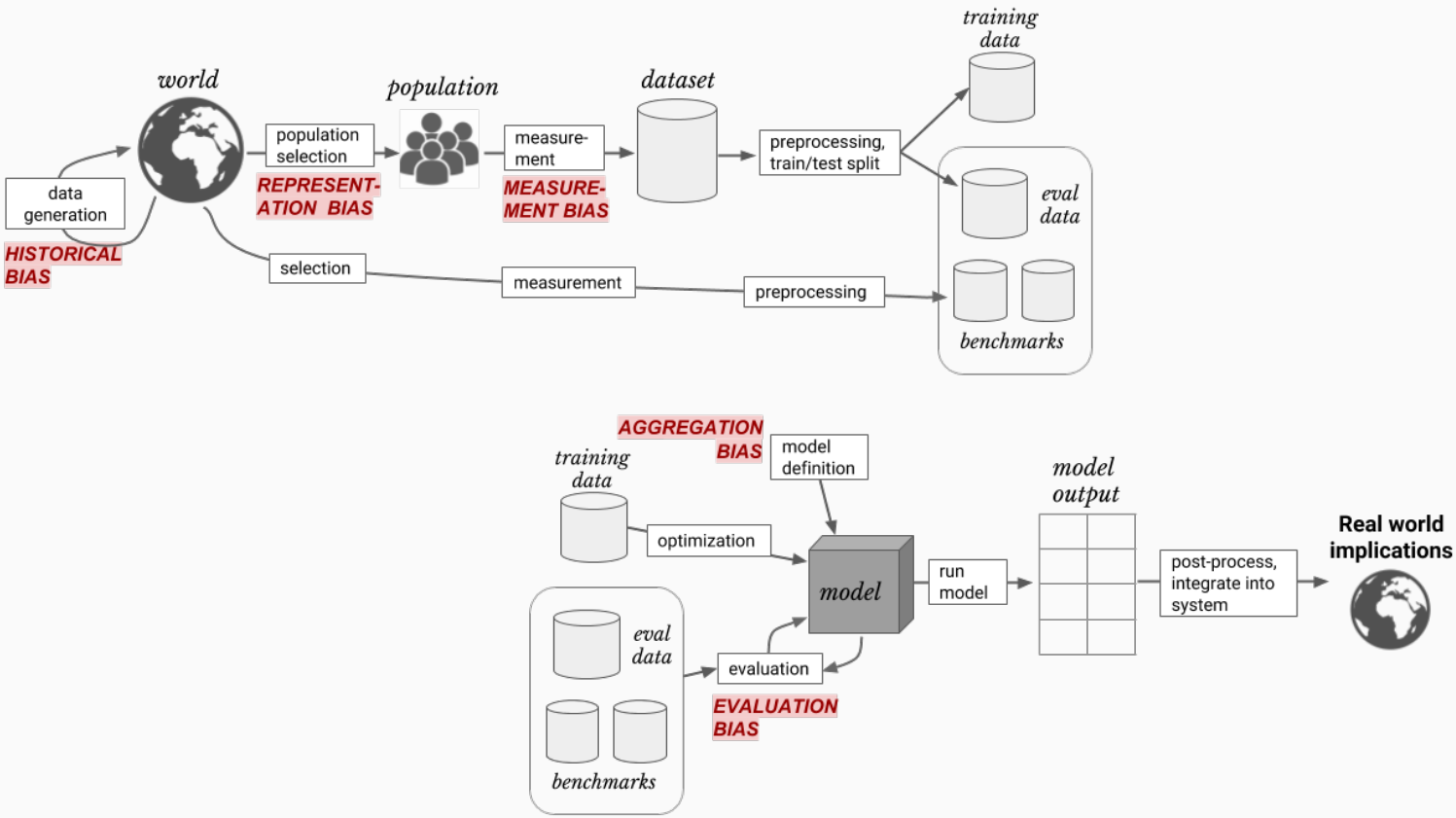


Figure 2: Fuentes de sesgo en el pipeline de ML[7]

[7] Harini Suresh and John Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle". In: *Equity and Access in Algorithms, Mechanisms, and Optimization.* – NY USA, 2021, pp. 1–9.

Fuentes de sesgo

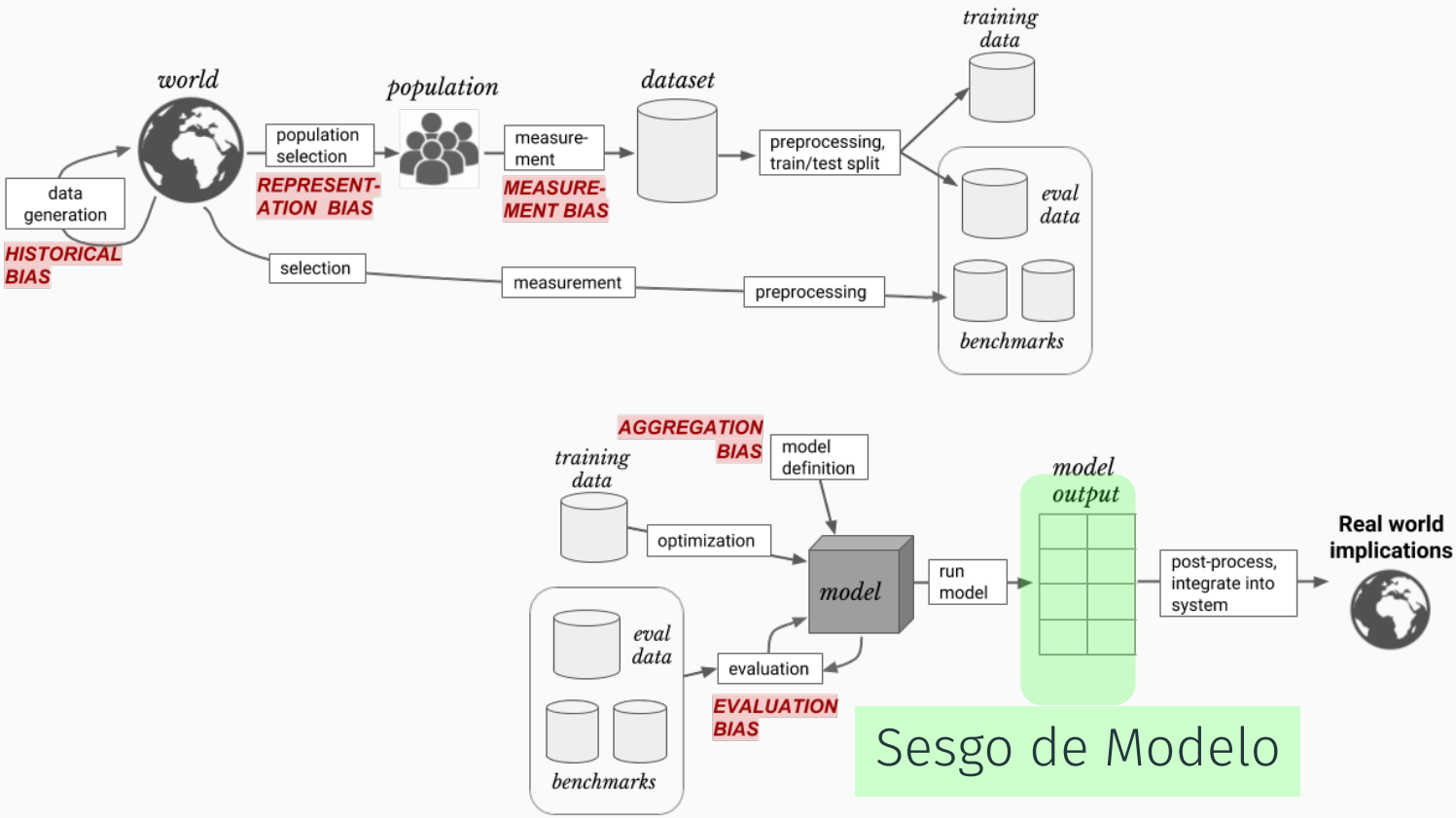


Figure 2: Fuentes de sesgo en el pipeline de ML[7]

[7] Harini Suresh and John Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle". In: *Equity and Access in Algorithms, Mechanisms, and Optimization.* – NY USA, 2021, pp. 1–9.

Fuentes de sesgo

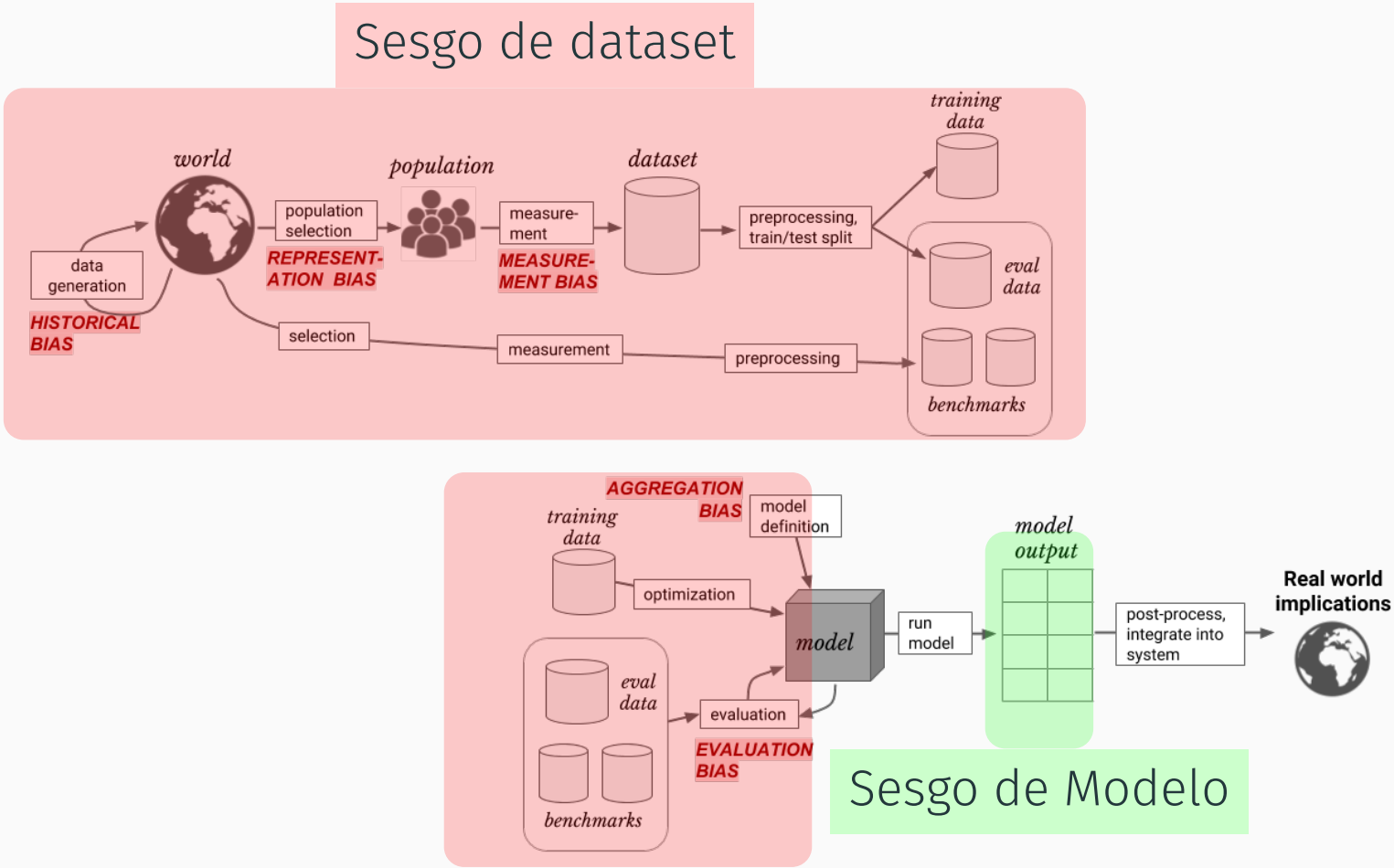


Figure 2: Fuentes de sesgo en el pipeline de ML[7]

[7] Harini Suresh and John Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle". In: *Equity and Access in Algorithms, Mechanisms, and Optimization.* – NY USA, 2021, pp. 1–9.

Fuentes de sesgo

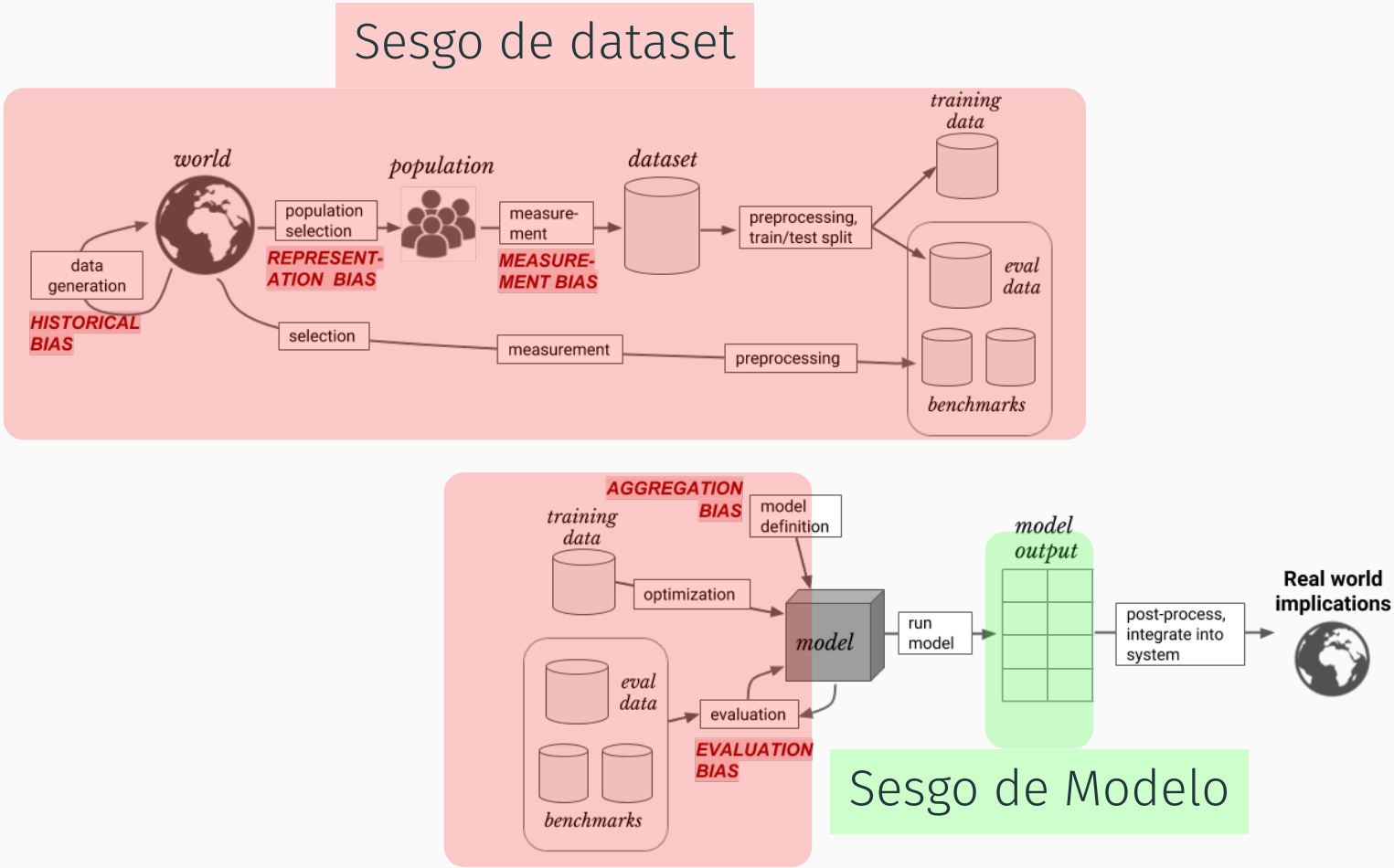


Figure 2: Fuentes de sesgo en el pipeline de ML[7]

[7] Harini Suresh and John Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle". In: *Equity and Access in Algorithms, Mechanisms, and Optimization.* – NY USA, 2021, pp. 1–9.

Sesgo en datasets, un ejemplo

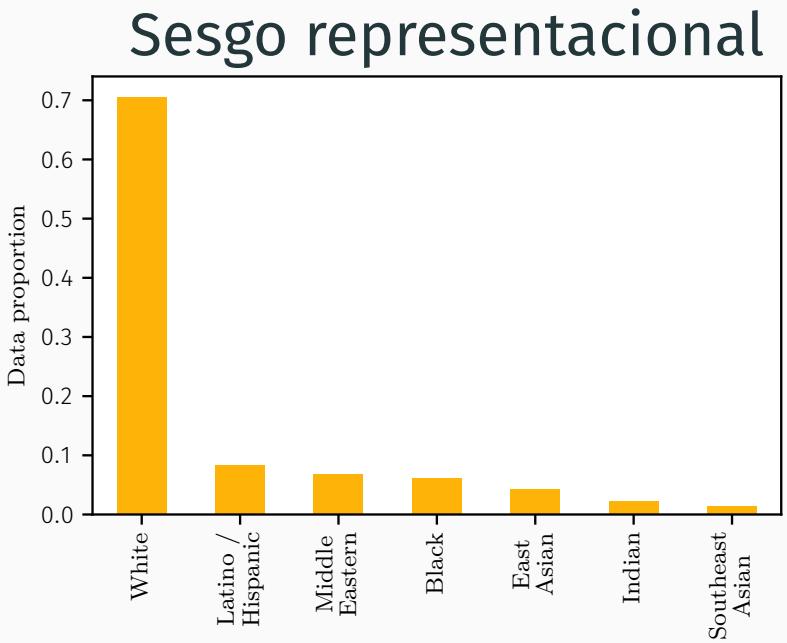


Figure 3: Distribución de raza aparente en FER+.

⚠ *No hay taxonomías o r... para este sesgo.*

⚠ *Ni estudios sobre su impacto.*

Sesgo en datasets, un ejemplo

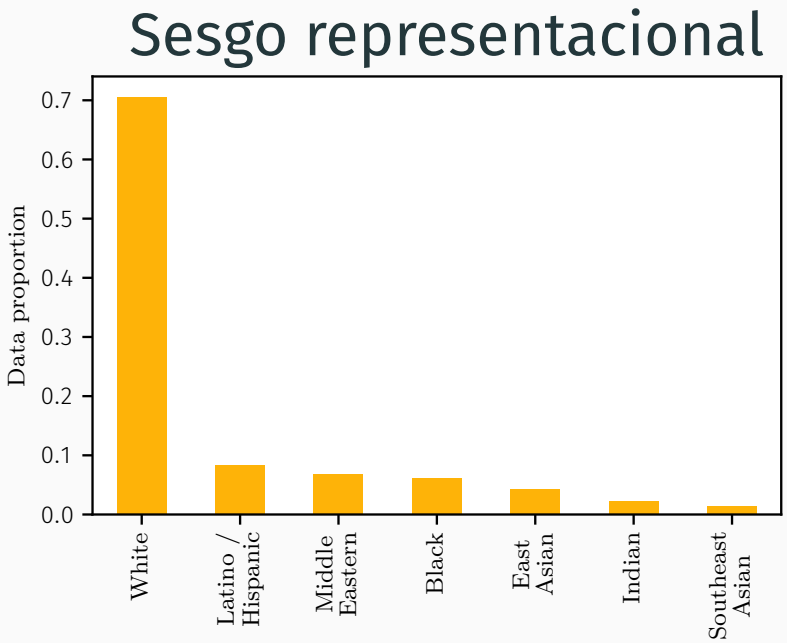


Figure 3: Distribución de raza aparente en FER+.

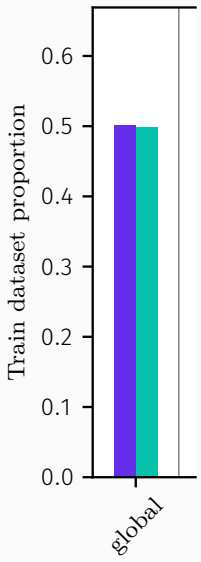


Figure 4: Distribución de género aparente *por clase* en FER+.

⚠ *No hay taxonomías o métricas para estos sesgos.*

⚠ *Ni estudios sobre su impacto.*

Sesgo en datasets, un ejemplo

Sesgo representacional

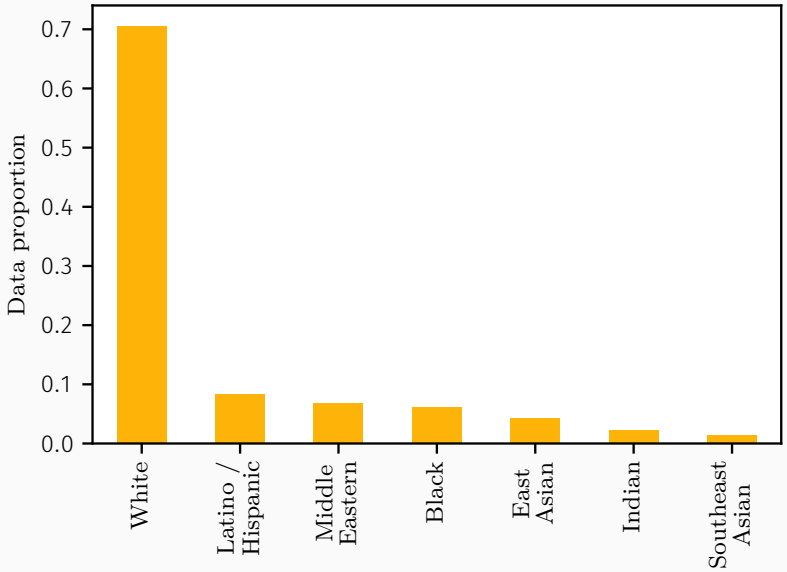


Figure 3: Distribución de raza aparente en FER+.

Sesgo estereotípico

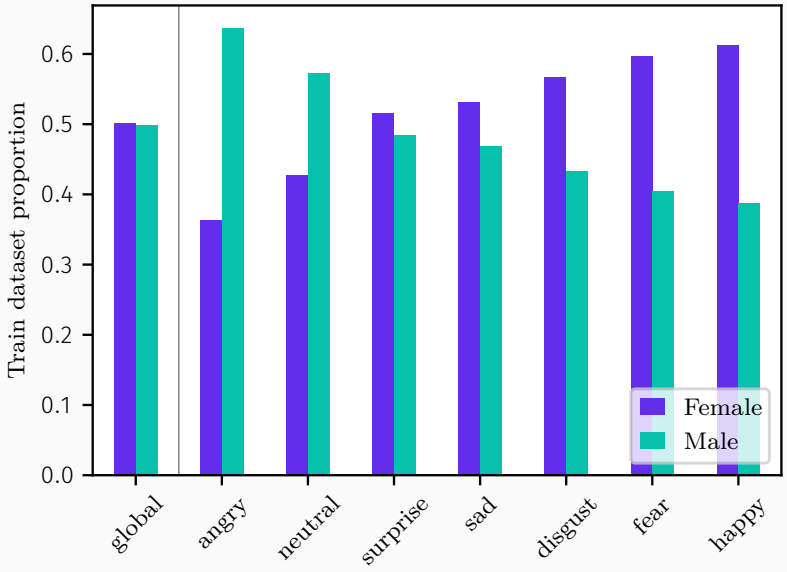


Figure 4: Distribución de género aparente *por clase* en FER+.

⚠ *No hay taxonomías o métricas para estos sesgos.*

⚠ *Ni estudios sobre su impacto.*

Sesgo en datasets, un ejemplo

Sesgo representacional

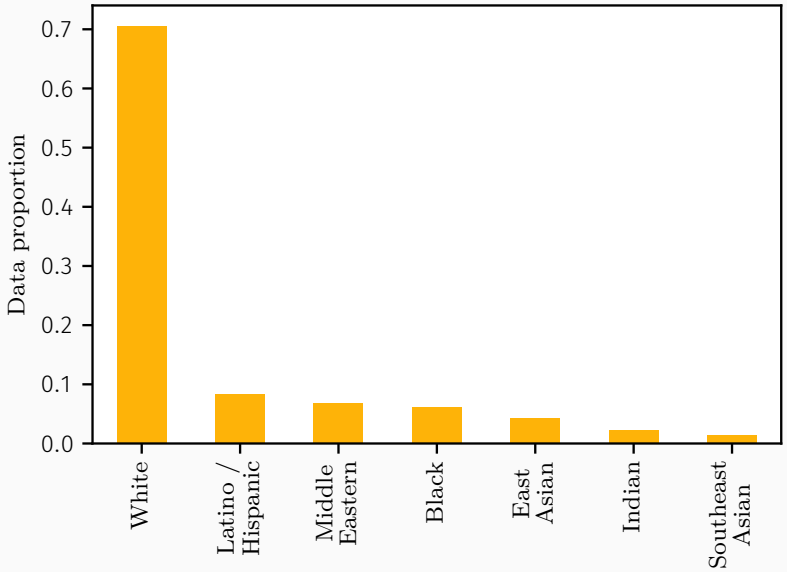


Figure 3: Distribución de raza aparente en FER+.

Sesgo estereotípico

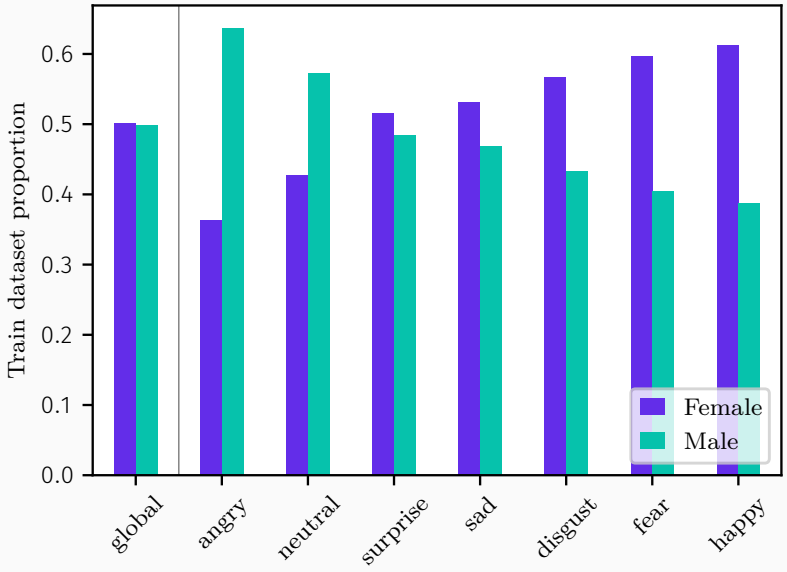


Figure 4: Distribución de género aparente *por clase* en FER+.

⚠ *No hay taxonomías o métricas para estos sesgos.*

⚠ *Ni estudios sobre su impacto.*

Introducción

Objetivos

Propuestas

Conclusiones y trabajo futuro

Objetivos

General: Investigar la transferencia del sesgo de los datasets a los modelos.

Específicos:

1. Desarrollar **una taxonomía del sesgo en datasets**, así como una revisión de sus métricas.
2. Crear **mejores métricas de sesgo en conjuntos de datos**, más interpretables y aplicables a datasets sin información demográfica.
3. **Comparar el impacto de los sesgos representacionales y estereotípicos** de los datasets en los modelos.
4. Analizar la **transferencia del sesgo** y desarrollar **métricas de sesgo en modelos** para problemas multiclase y multigrupo.

Objetivos

General: Investigar la transferencia del sesgo de los datasets a los modelos.

Específicos:

1. Desarrollar **una taxonomía del sesgo en datasets**, así como una revisión de sus métricas.
2. Crear **mejores métricas de sesgo en conjuntos de datos**, más interpretables y aplicables a datasets sin información demográfica.
3. **Comparar el impacto de los sesgos representacionales y estereotípicos** de los datasets en los modelos.
4. Analizar la **transferencia del sesgo** y desarrollar **métricas de sesgo en modelos** para problemas multiclase y multigrupo.

Objetivos

General: Investigar la transferencia del sesgo de los datasets a los modelos.

Específicos:

1. Desarrollar **una taxonomía del sesgo en datasets**, así como una revisión de sus métricas.
2. Crear **mejores métricas de sesgo en conjuntos de datos**, más interpretables y aplicables a datasets sin información demográfica.
3. Comparar el impacto de los sesgos representacionales y **estereotípicos** de los datasets en los modelos.
4. Analizar la **transferencia del sesgo** y desarrollar **métricas de sesgo en modelos** para problemas multiclase y multigrupo.

Objetivos

General: Investigar la transferencia del sesgo de los datasets a los modelos.

Específicos:

1. Desarrollar **una taxonomía del sesgo en datasets**, así como una revisión de sus métricas.
2. Crear **mejores métricas de sesgo en conjuntos de datos**, más interpretables y aplicables a datasets sin información demográfica.
3. **Comparar el impacto de los sesgos representacionales y estereotípicos** de los datasets en los modelos.
4. Analizar la **transferencia del sesgo** y desarrollar **métricas de sesgo en modelos** para problemas multiclase y multigrupo.

Objetivos

General: Investigar la transferencia del sesgo de los datasets a los modelos.

Específicos:

1. Desarrollar **una taxonomía del sesgo en datasets**, así como una revisión de sus métricas.
2. Crear **mejores métricas de sesgo en conjuntos de datos**, más interpretables y aplicables a datasets sin información demográfica.
3. **Comparar el impacto de los sesgos representacionales y estereotípicos** de los datasets en los modelos.
4. Analizar la **transferencia del sesgo** y desarrollar **métricas de sesgo en modelos** para problemas multiclase y multigrupo.

Trabajos

- **Artículo 1**: Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. “Metrics for Dataset Demographic Bias: A Case Study on Facial Expression Recognition”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024), pp. 1–18
- **Artículo 2**: Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. “DSAP: Analyzing Bias through Demographic Comparison of Datasets”. In: *Information Fusion* 115 (2025), p. 102760
- **Artículo 3**: Iris Dominguez-Catena et al. “Less Can Be More: Representational vs. Stereotypical Gender Bias in Facial Expression Recognition”. In: *Prog Artif Intell* (2024)
- **Artículo 4**: Iris Dominguez-Catena et al. *Biased Heritage: How Datasets Shape Models in Facial Expression Recognition*. 2025. arXiv: `2503.03446 [cs]` (En revisión en IEEE Transactions on Affective Computing)
- **IJCAI-ECAI 2022**: Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. “Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition”. In: *Proceedings of the Workshop on Artificial Intelligence Safety 2022 (AISafety 2022)*. Vienna, Austria, 2022
- **ECML-PKDD 2022**: Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. *Metrics for Dataset Demographic Bias: A Case Study on Facial Expression Recognition*. 2023. arXiv: `2303.15889 [cs]`

Introducción

Objetivos

Propuestas

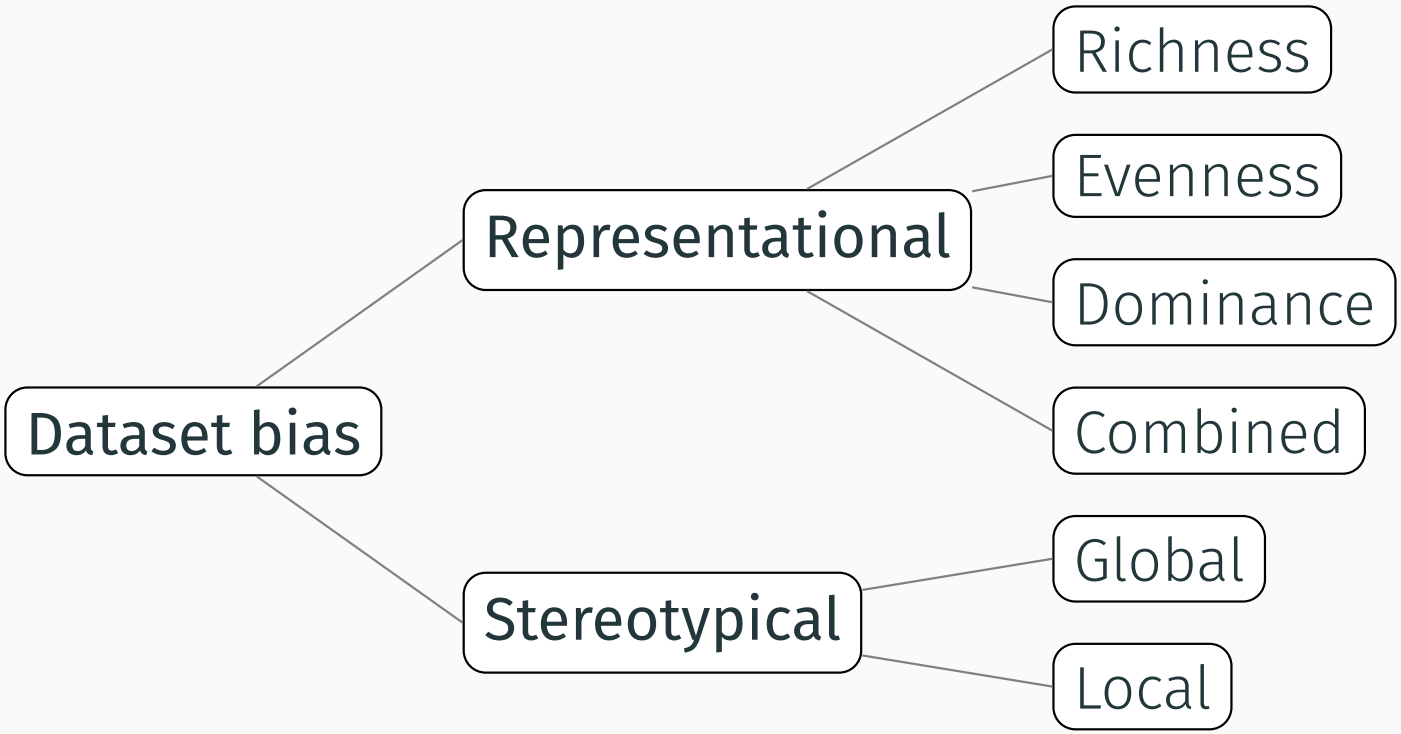
- Metrics for Dataset Demographic Bias
- Analyzing Bias Through Demographic Comparison of Datasets
- Representational vs. Stereotypical Bias Transference
- Measuring Transference from Dataset Bias to Model Predictions

Conclusiones y trabajo futuro

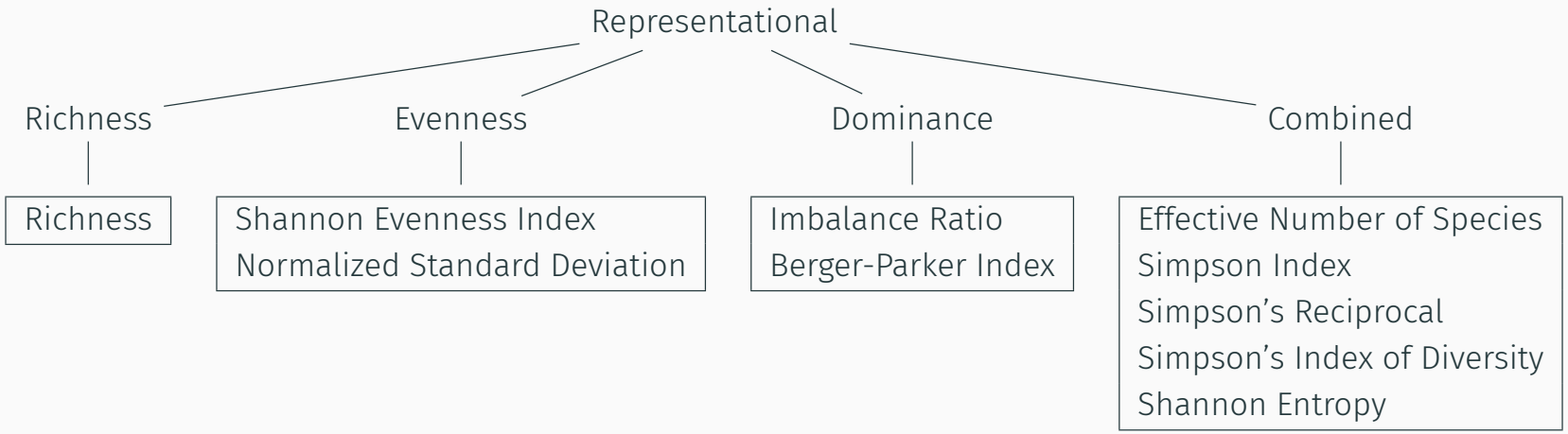
Propuestas

Metrics for Dataset Demographic Bias

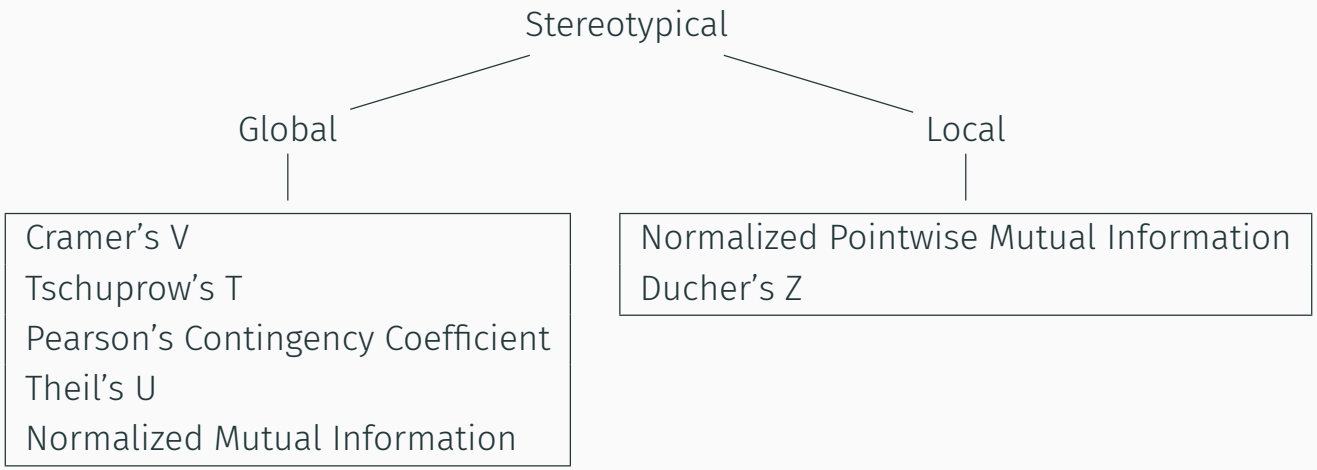
Dataset bias taxonomy



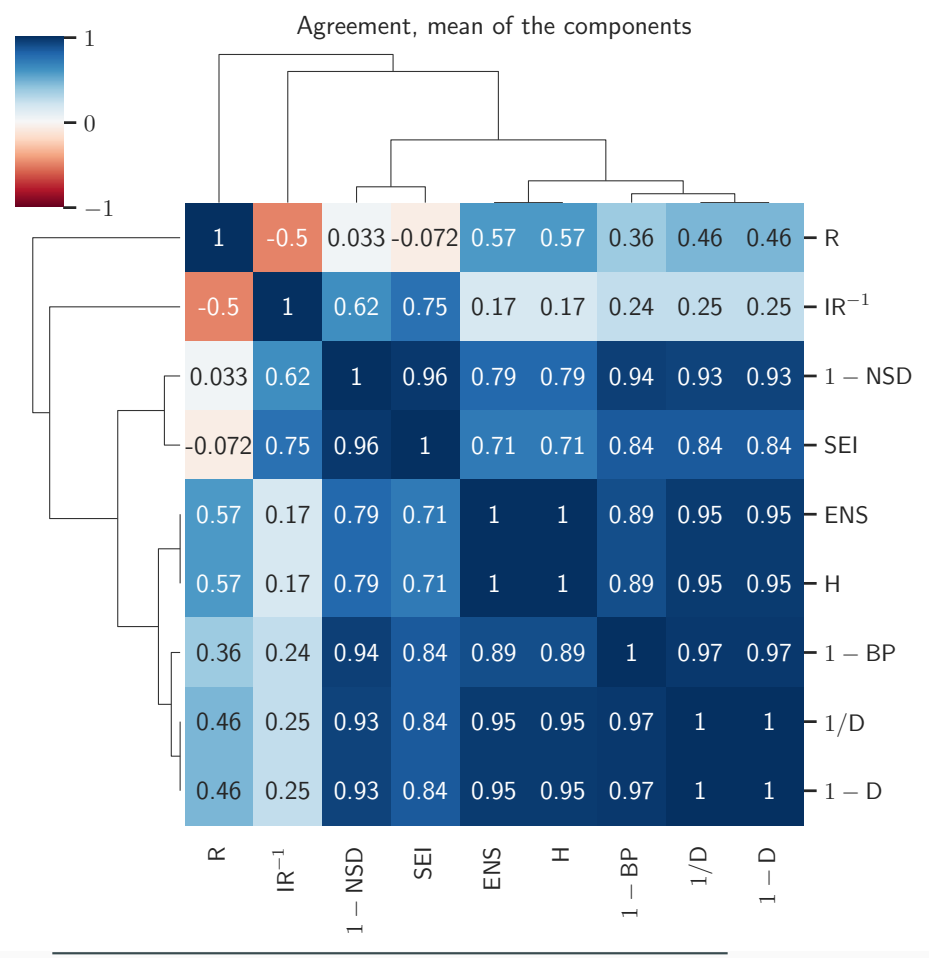
Representational bias metrics



Stereotypical bias metrics



Coherencia entre métricas (represent.)



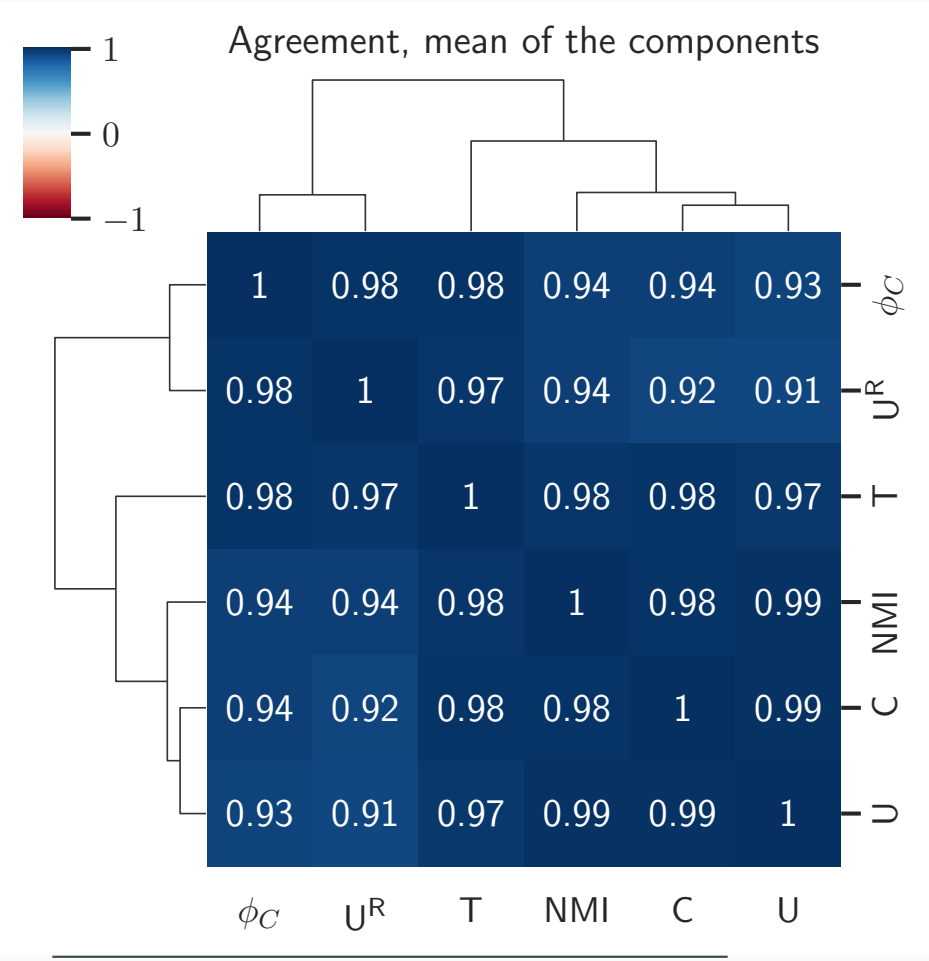
- General: **Effective Number of Species** (ENS) [8]
- Evenness / Homogeneidad: **Shannon Evenness Index** (SEI) [9]
- Dominancia: **Berger-Parker Index** (BP) [10]

[8] Lou Jost. "Entropy and Diversity". In: *Oikos* 113.2 (2006), pp. 363–375

[9] E.C. Pielou. "The Measurement of Diversity in Different Types of Biological Collections". In: *Journal of Theoretical Biology* 13 (1966), pp. 131–144

[10] Wolfgang H. Berger and Frances L. Parker. "Diversity of Planktonic Foraminifera in Deep-Sea Sediments". In: *Science* 168.3937 (1970), pp. 1345–1347

Coherencia entre métricas (estereo.)



- Global: **Cramer's V** (ϕ_C) [11]
- Local: **Ducher's Z** (Z) [12]

[11] Harald Cramér. "Chapter 21. The Two-Dimensional Case". In: *Mathematical Methods of Statistics*. Princeton Mathematical Series 9. Princeton, 1991, p. 282

[12] M. Ducher et al. "Statistical Relationships between Systolic Blood Pressure and Heart Rate and Their Functional Significance in Conscious Rats". In: *Med. Biol. Eng. Comput.* 32.6 (1994), pp. 649–655

		Laboratory	ITW-M	ITW-I
Age	$9-$ ENS	7.067 ± 0.932	5.551 ± 0.990	3.334 ± 0.286
	$1-$ SEI	0.409 ± 0.200	0.437 ± 0.154	0.211 ± 0.024
	ϕ_C	0.075 ± 0.063	0.084 ± 0.031	0.104 ± 0.027
Race	$7-$ ENS	5.168 ± 0.634	5.070 ± 0.080	3.724 ± 0.321
	$1-$ SEI	0.384 ± 0.151	0.663 ± 0.021	0.393 ± 0.050
	ϕ_C	0.092 ± 0.083	0.058 ± 0.018	0.063 ± 0.018
Gender	$2-$ ENS	0.139 ± 0.280	0.074 ± 0.055	0.005 ± 0.005
	$1-$ SEI	0.039 ± 0.052	0.055 ± 0.041	0.004 ± 0.004
	ϕ_C	0.067 ± 0.090	0.199 ± 0.062	0.167 ± 0.018

Figure 5: Sesgo representacional medio (ENS), homogeneidad (SEI) y sesgo estereotípico (ϕ_C) de datasets Lab, ITW-M y ITW-I.

		Laboratory	ITW-M	ITW-I
Age	$9-$ ENS	7.067 ± 0.932	5.551 ± 0.990	3.334 ± 0.286
	$1-$ SEI	0.409 ± 0.200	0.437 ± 0.154	0.211 ± 0.024
	ϕ_C	0.075 ± 0.063	0.084 ± 0.031	0.104 ± 0.027
Race	$7-$ ENS	5.168 ± 0.634	5.070 ± 0.080	3.724 ± 0.321
	$1-$ SEI	0.384 ± 0.151	0.663 ± 0.021	0.393 ± 0.050
	ϕ_C	0.092 ± 0.083	0.058 ± 0.018	0.063 ± 0.018
Gender	$2-$ ENS	0.139 ± 0.280	0.074 ± 0.055	0.005 ± 0.005
	$1-$ SEI	0.039 ± 0.052	0.055 ± 0.041	0.004 ± 0.004
	ϕ_C	0.067 ± 0.090	0.199 ± 0.062	0.167 ± 0.018

Figure 5: Sesgo representacional medio (ENS), homogeneidad (SEI) y sesgo estereotípico (ϕ_C) de datasets Lab, ITW-M y ITW-I.

		Laboratory	ITW-M	ITW-I
Age	$9-$ ENS	7.067 ± 0.932	5.551 ± 0.990	3.334 ± 0.286
	$1-$ SEI	0.409 ± 0.200	0.437 ± 0.154	0.211 ± 0.024
	ϕ_C	0.075 ± 0.063	0.084 ± 0.031	0.104 ± 0.027
Race	$7-$ ENS	5.168 ± 0.634	5.070 ± 0.080	3.724 ± 0.321
	$1-$ SEI	0.384 ± 0.151	0.663 ± 0.021	0.393 ± 0.050
	ϕ_C	0.092 ± 0.083	0.058 ± 0.018	0.063 ± 0.018
Gender	$2-$ ENS	0.139 ± 0.280	0.074 ± 0.055	0.005 ± 0.005
	$1-$ SEI	0.039 ± 0.052	0.055 ± 0.041	0.004 ± 0.004
	ϕ_C	0.067 ± 0.090	0.199 ± 0.062	0.167 ± 0.018

Figure 5: Sesgo representacional medio (ENS), homogeneidad (SEI) y sesgo estereotípico (ϕ_C) de datasets Lab, ITW-M y ITW-I.

Sesgo estereotípico local

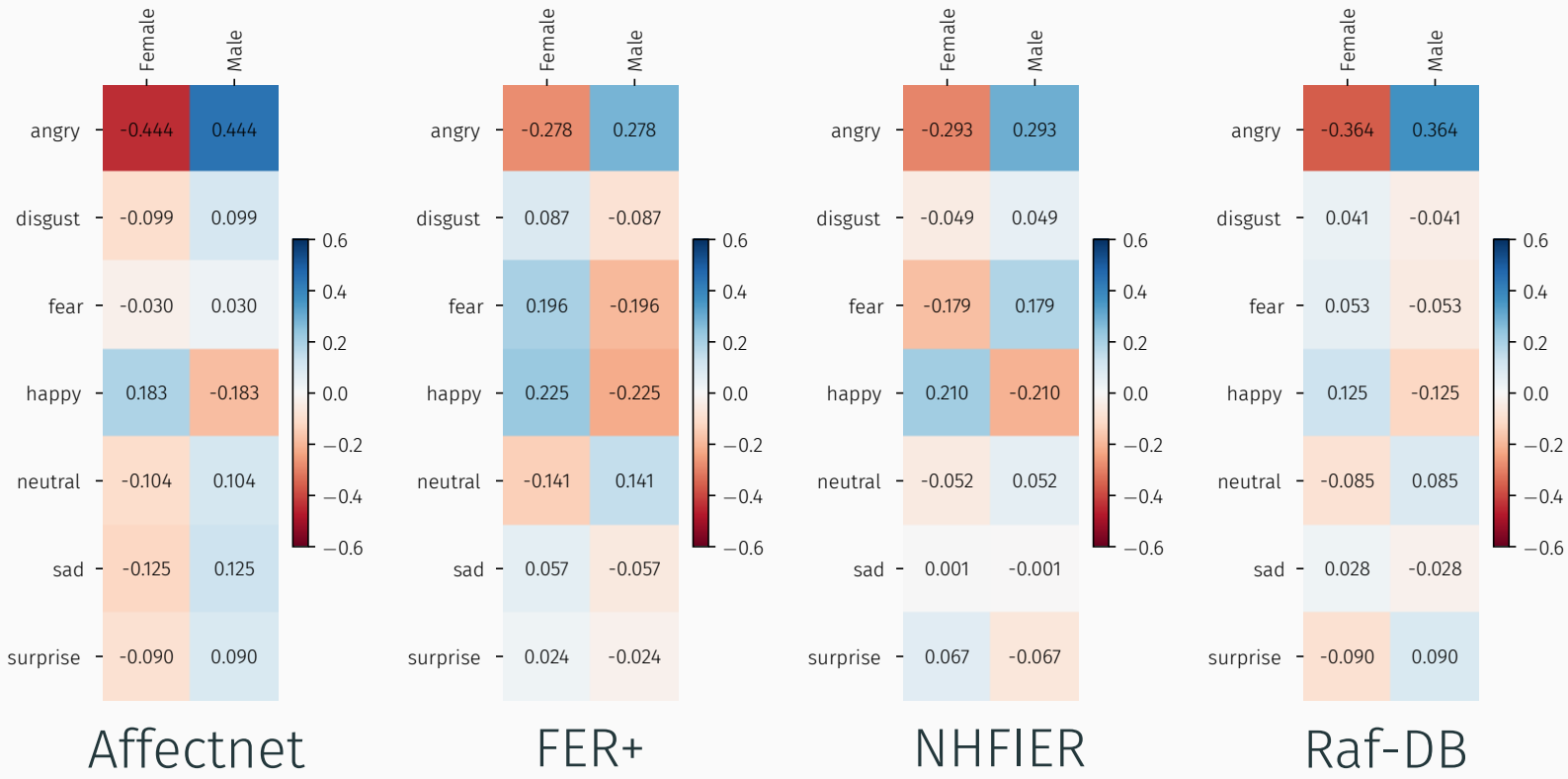


Figure 6: Sesgo estereotípico local (Ducher's Z) de algunos datasets ITW-I.

Sesgo estereotípico local

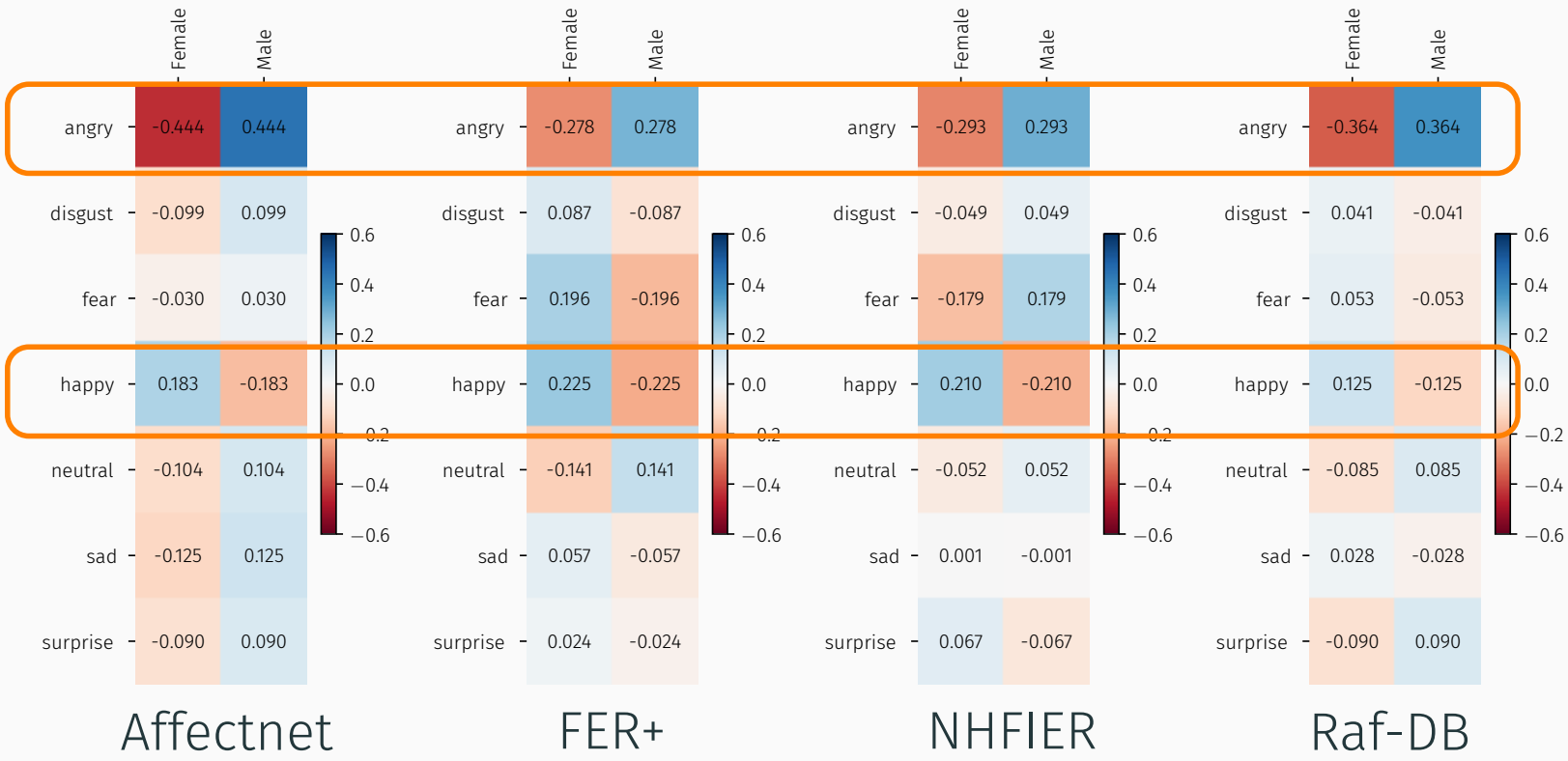


Figure 6: Sesgo estereotípico local (Ducher's Z) de algunos datasets ITW-I.

Propuestas

Analyzing Bias Through Demographic Comparison of Datasets

- Las métricas anteriores son poco **interpretables**.
  - Diferencias en los rangos.
  - Falta de significado.
- La mayor parte de datasets no tienen información demográfica.
- ¿Cómo estudiar la evolución del sesgo en los datasets?
  - Problemas análogos en arqueología[13] y ecología[14,15].

[13] W. S. Robinson. "A Method for Chronologically Ordering Archaeological Deposits". In: *Am. antiq.* 16.4 (1951), pp. 293–301.

[14] M. V. Wilson and A. Shmida. "Measuring Beta Diversity with Presence-Absence Data". In: *Journal of Ecology* 72.3 (1984), pp. 1055–1064. JSTOR: 2259551.

[15] C. Ricotta and J. Podani. "On Some Properties of the Bray-Curtis Dissimilarity and Their Ecological Meaning". In: *Ecological Complexity* 31 (2017), pp. 201–205.

- Las métricas anteriores son poco **interpretables**.
  - Diferencias en los rangos.
  - Falta de significado.
- La mayor parte de datasets no tienen información demográfica.
- ¿Cómo estudiar la evolución del sesgo en los datasets?
  - Problemas análogos en arqueología[13] y ecología[14,15].

[13] W. S. Robinson. "A Method for Chronologically Ordering Archaeological Deposits". In: *Am. antiq.* 16.4 (1951), pp. 293–301.

[14] M. V. Wilson and A. Shmida. "Measuring Beta Diversity with Presence-Absence Data". In: *Journal of Ecology* 72.3 (1984), pp. 1055–1064. JSTOR: 2259551.

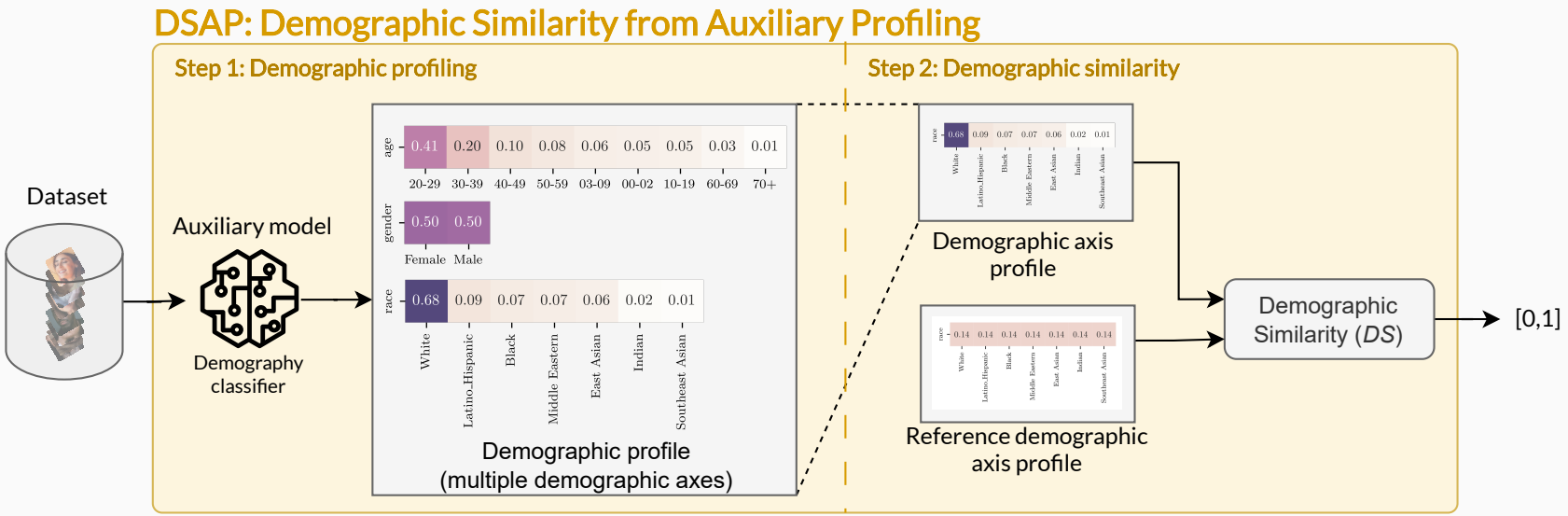
[15] C. Ricotta and J. Podani. "On Some Properties of the Bray-Curtis Dissimilarity and Their Ecological Meaning". In: *Ecological Complexity* 31 (2017), pp. 201–205.

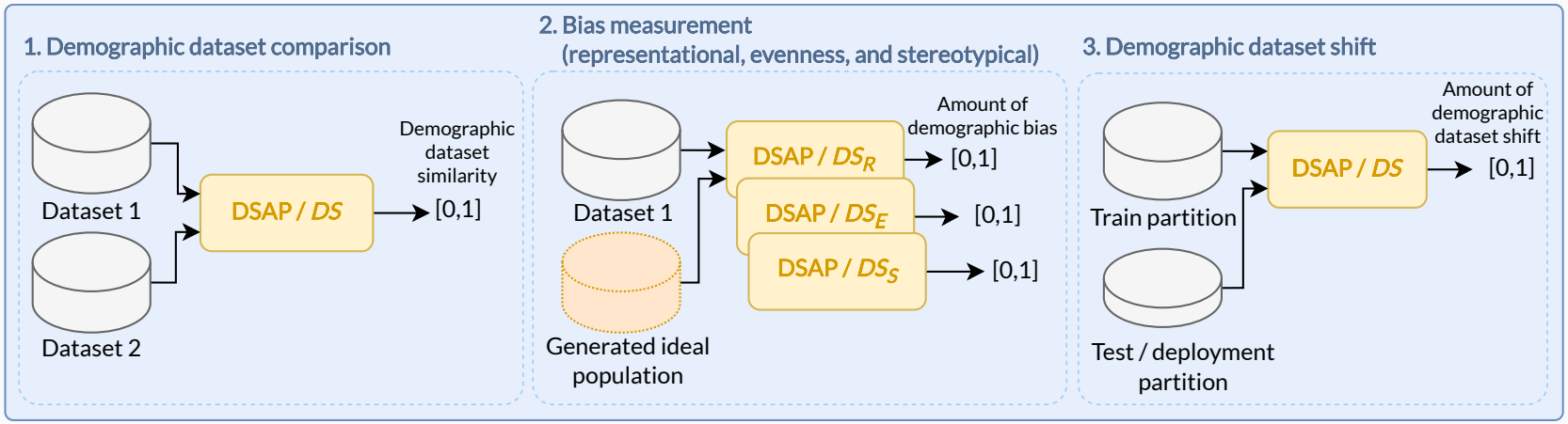
- Las métricas anteriores son poco **interpretables**.
  - Diferencias en los rangos.
  - Falta de significado.
- La mayor parte de datasets no tienen información demográfica.
- ¿Cómo estudiar la evolución del sesgo en los datasets?
  - Problemas análogos en arqueología[13] y ecología[14,15].

[13] W. S. Robinson. "A Method for Chronologically Ordering Archaeological Deposits". In: *Am. antiq.* 16.4 (1951), pp. 293–301.

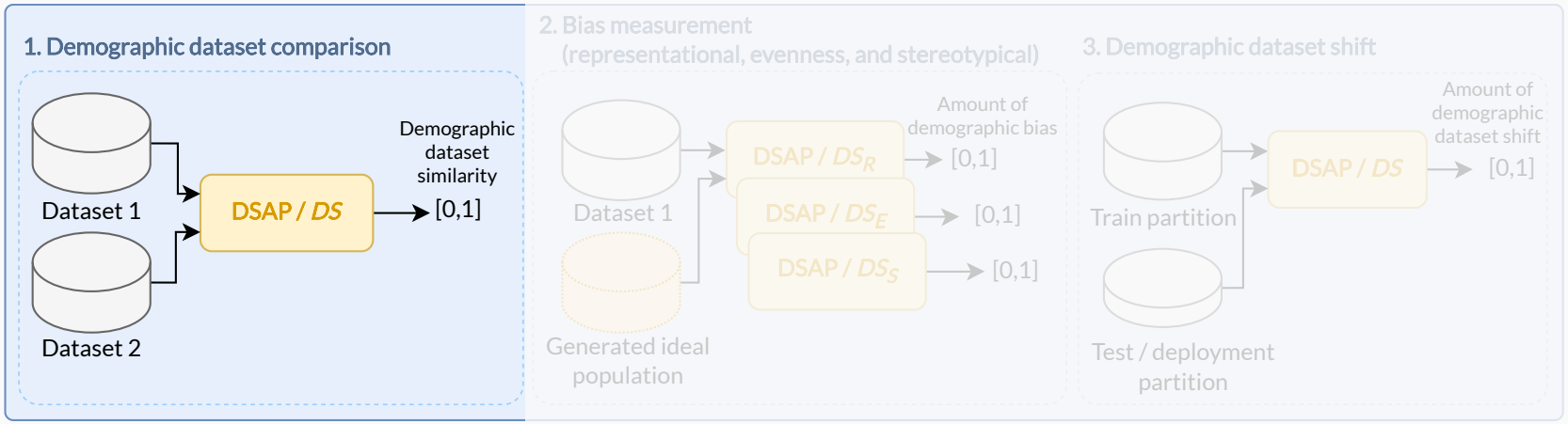
[14] M. V. Wilson and A. Shmida. "Measuring Beta Diversity with Presence-Absence Data". In: *Journal of Ecology* 72.3 (1984), pp. 1055–1064. JSTOR: 2259551.

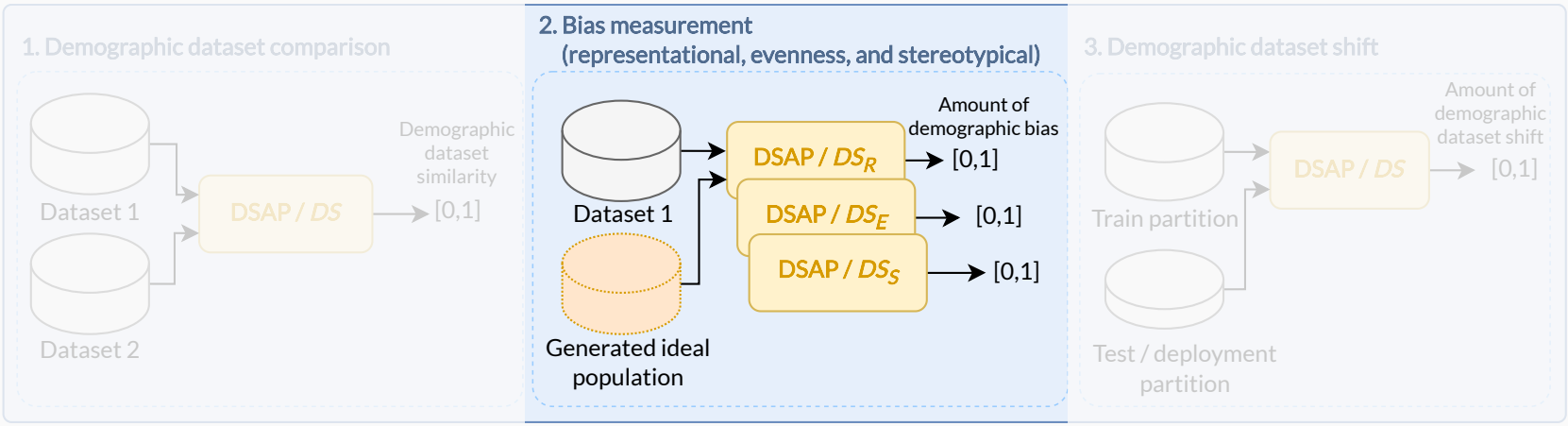
[15] C. Ricotta and J. Podani. "On Some Properties of the Bray-Curtis Dissimilarity and Their Ecological Meaning". In: *Ecological Complexity* 31 (2017), pp. 201–205.

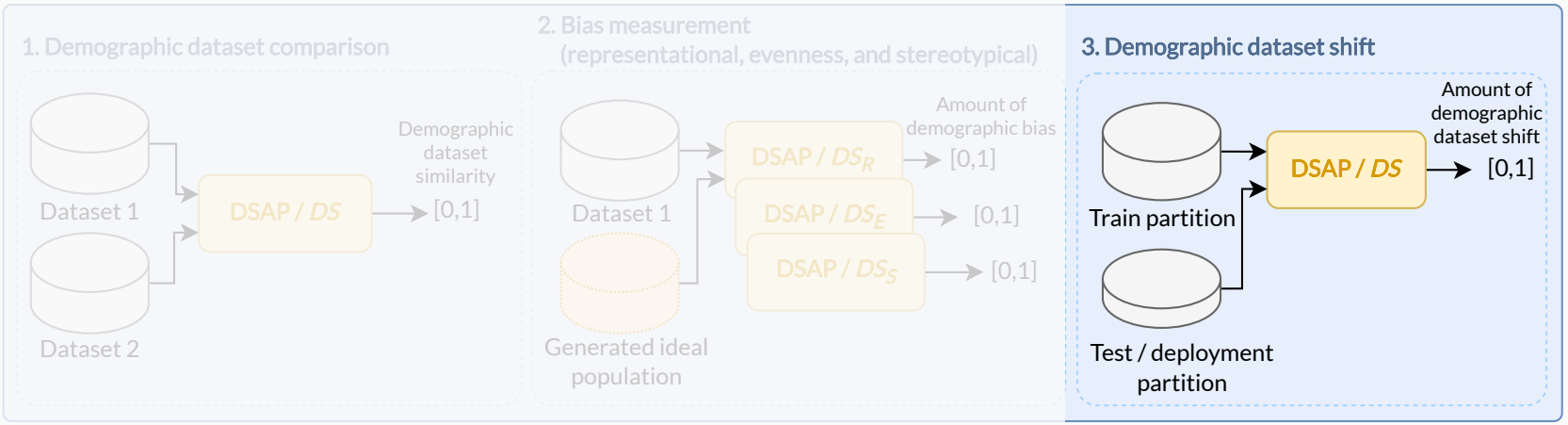

DSAP: Demographic Similarity from Auxiliary Profiling
Step 1: Demographic profiling
Step 2: Demographic similarity
Dataset
Auxiliary model
Demography classifier
age
0.41 0.20 0.10 0.08 0.06 0.05 0.05 0.03 0.01
20-29 30-39 40-49 50-59 03-09 00-02 10-19 60-69 70+
gender
0.50 0.50
Female Male
race
0.68 0.09 0.07 0.07 0.06 0.02 0.01
White Latino_Hispanic Black Middle Eastern East Asian Indian Southeast Asian
Demographic profile
(multiple demographic axes)
Demographic axis profile
Reference demographic axis profile
Demographic Similarity (DS)
[0,1]


1. Demographic dataset comparison
Dataset 1
Dataset 2
DSAP / DS
Demographic dataset similarity
[0,1]
2. Bias measurement (representational, evenness, and stereotypical)
Dataset 1
Generated ideal population
DSAP / DS_R
DSAP / DS_E
DSAP / DS_S
Amount of demographic bias
[0,1]
[0,1]
[0,1]
3. Demographic dataset shift
Train partition
Test / deployment partition
DSAP / DS
Amount of demographic dataset shift
[0,1]

Applications




1. Demographic dataset comparison
Dataset 1
Dataset 2
DSAP / DS
Demographic dataset similarity
[0,1]
2. Bias measurement (representational, evenness, and stereotypical)
Dataset 1
Generated ideal population
DSAP / DS_R
DSAP / DS_E
DSAP / DS_S
Amount of demographic bias
[0,1]
[0,1]
[0,1]
3. Demographic dataset shift
Train partition
Test / deployment partition
DSAP / DS
Amount of demographic dataset shift
[0,1]


1. Demographic dataset comparison
Dataset 1
Dataset 2
DSAP / DS
Demographic dataset similarity
[0,1]
2. Bias measurement (representational, evenness, and stereotypical)
Dataset 1
Generated ideal population
DSAP / DS_R
DSAP / DS_E
DSAP / DS_S
Amount of demographic bias
[0,1]
[0,1]
[0,1]
3. Demographic dataset shift
Train partition
Test / deployment partition
DSAP / DS
Amount of demographic dataset shift
[0,1]

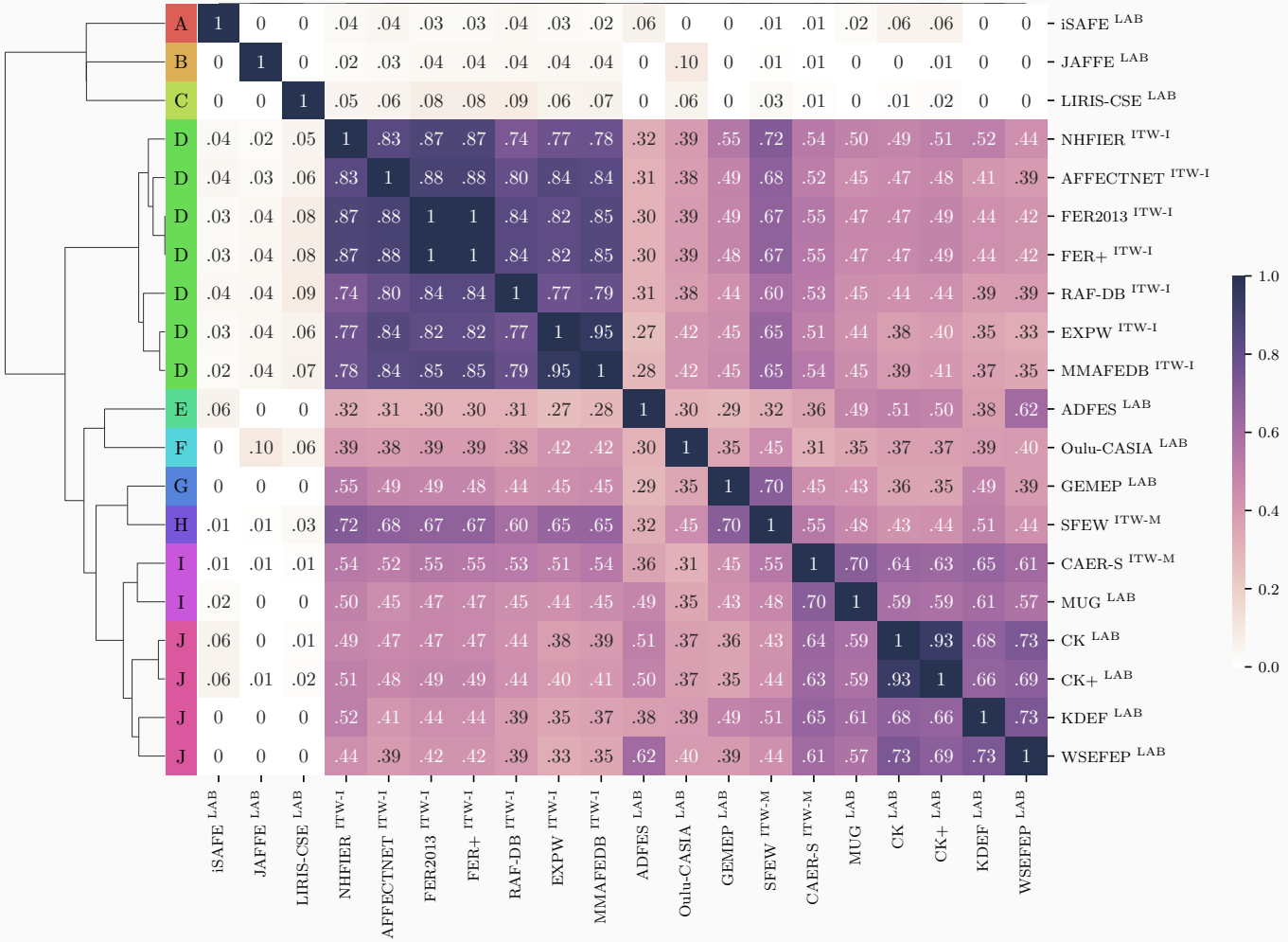


Figure 7: Comparación de datasets basada en DSAP (interseccional, 126 subgrupos).

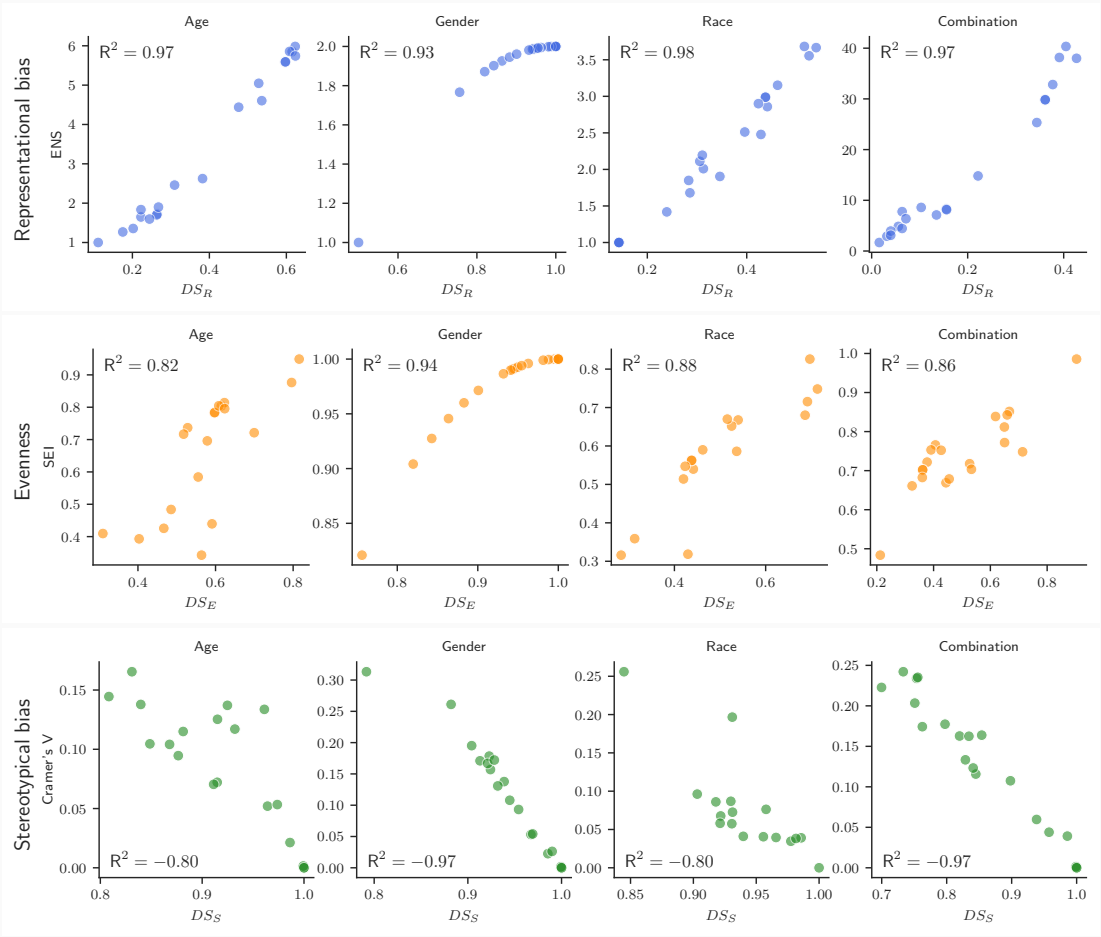


Figure 8: Métricas basadas en DSAP (eje x) comparadas con sus contrapartidas anteriores (eje y).

Propuestas

Representational vs. Stereotypical Bias Transference

Hemos visto que el sesgo estereotípico se ha vuelto prevalente en los datasets ITW. ¿Qué efecto tiene esto?

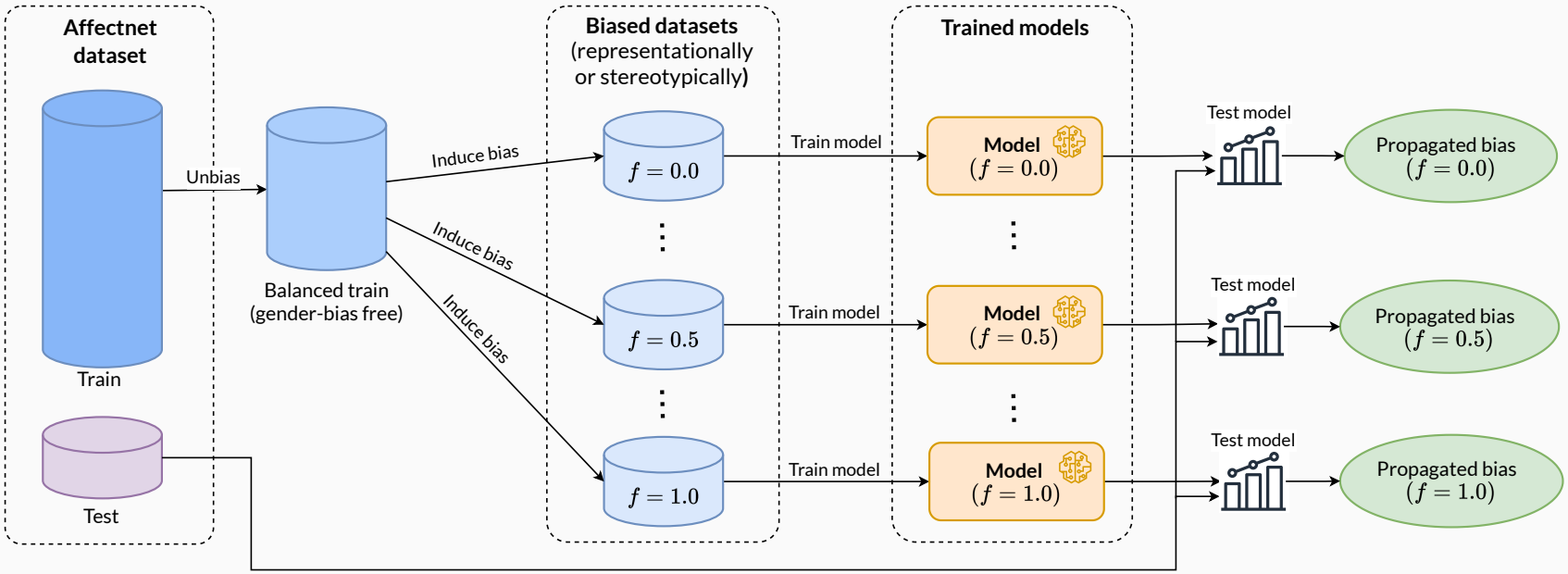


Figure 9: Resumen de la metodología

Inducción de sesgo

Original

Gender	Female	Male	% F
angry	6,962	17,803	28.11%
disgust	1,733	2,054	45.76%
fear	3,117	3,222	49.17%
happy	79,797	54,241	59.53%
neutral	33,708	40,858	45.21%
sad	11,175	14,156	44.12%
surprise	6,439	7,588	45.90%
Total	142,931	139,922	50.53%

Balanceado

Gender	Female	Male	% F
angry	6,962	6,962	50%
disgust	1,733	1,733	50%
fear	3,117	3,117	50%
happy	54,241	54,241	50%
neutral	33,708	33,708	50%
sad	11,175	11,175	50%
surprise	6,439	6,439	50%
Total	117,375	117,375	50%

Sesgo representacional

Gender	Female	Male	% F
angry	6,962	0	100%
disgust	1,733	0	100%
fear	3,117	0	100%
happy	54,241	0	100%
neutral	33,708	0	100%
sad	11,175	0	100%
surprise	6,439	0	100%
Total	117,375	0	100%

Sesgo estereotípico

Gender	Female	Male	% F
angry	0	6,962	0%
disgust	866	866	50%
fear	1,558	1,558	50%
happy	27,120	27,120	50%
neutral	16,854	16,854	50%
sad	5,587	5,587	50%
surprise	3,219	3,219	50%
Total	55,204	62,166	47.4%

Inducción de sesgo

Original

Gender	Female	Male	% F
angry	6,962	17,803	28.11%
disgust	1,733	2,054	45.76%
fear	3,117	3,222	49.17%
happy	79,797	54,241	59.53%
neutral	33,708	40,858	45.21%
sad	11,175	14,156	44.12%
surprise	6,439	7,588	45.90%
Total	142,931	139,922	50.53%

Balanceado

Gender	Female	Male	% F
angry	6,962	6,962	50%
disgust	1,733	1,733	50%
fear	3,117	3,117	50%
happy	54,241	54,241	50%
neutral	33,708	33,708	50%
sad	11,175	11,175	50%
surprise	6,439	6,439	50%
Total	117,375	117,375	50%

Sesgo representacional

Gender	Female	Male	% F
angry	6,962	0	100%
disgust	1,733	0	100%
fear	3,117	0	100%
happy	54,241	0	100%
neutral	33,708	0	100%
sad	11,175	0	100%
surprise	6,439	0	100%
Total	117,375	0	100%

Sesgo estereotípico

Gender	Female	Male	% F
angry	0	6,962	0%
disgust	866	866	50%
fear	1,558	1,558	50%
happy	27,120	27,120	50%
neutral	16,854	16,854	50%
sad	5,587	5,587	50%
surprise	3,219	3,219	50%
Total	55,204	62,166	47.4%

Inducción de sesgo

Original

Gender	Female	Male	% F
angry	6,962	17,803	28.11%
disgust	1,733	2,054	45.76%
fear	3,117	3,222	49.17%
happy	79,797	54,241	59.53%
neutral	33,708	40,858	45.21%
sad	11,175	14,156	44.12%
surprise	6,439	7,588	45.90%
Total	142,931	139,922	50.53%

Balanceado

Gender	Female	Male	% F
angry	6,962	6,962	50%
disgust	1,733	1,733	50%
fear	3,117	3,117	50%
happy	54,241	54,241	50%
neutral	33,708	33,708	50%
sad	11,175	11,175	50%
surprise	6,439	6,439	50%
Total	117,375	117,375	50%

Sesgo representacional

Gender	Female	Male	% F
angry	6,962	0	100%
disgust	1,733	0	100%
fear	3,117	0	100%
happy	54,241	0	100%
neutral	33,708	0	100%
sad	11,175	0	100%
surprise	6,439	0	100%
Total	117,375	0	100%

Sesgo estereotípico

Gender	Female	Male	% F
angry	0	6,962	0%
disgust	866	866	50%
fear	1,558	1,558	50%
happy	27,120	27,120	50%
neutral	16,854	16,854	50%
sad	5,587	5,587	50%
surprise	3,219	3,219	50%
Total	55,204	62,166	47.4%

Inducción de sesgo

Original

Gender	Female	Male	% F
angry	6,962	17,803	28.11%
disgust	1,733	2,054	45.76%
fear	3,117	3,222	49.17%
happy	79,797	54,241	59.53%
neutral	33,708	40,858	45.21%
sad	11,175	14,156	44.12%
surprise	6,439	7,588	45.90%
Total	142,931	139,922	50.53%

Balanceado

Gender	Female	Male	% F
angry	6,962	6,962	50%
disgust	1,733	1,733	50%
fear	3,117	3,117	50%
happy	54,241	54,241	50%
neutral	33,708	33,708	50%
sad	11,175	11,175	50%
surprise	6,439	6,439	50%
Total	117,375	117,375	50%

Sesgo representacional

Gender	Female	Male	% F
angry	6,962	0	100%
disgust	1,733	0	100%
fear	3,117	0	100%
happy	54,241	0	100%
neutral	33,708	0	100%
sad	11,175	0	100%
surprise	6,439	0	100%
Total	117,375	0	100%

Sesgo estereotípico

Gender	Female	Male	% F
angry	0	6,962	0%
disgust	866	866	50%
fear	1,558	1,558	50%
happy	27,120	27,120	50%
neutral	16,854	16,854	50%
sad	5,587	5,587	50%
surprise	3,219	3,219	50%
Total	55,204	62,166	47.4%

Inducción de sesgo

Original

Gender	Female	Male	% F
angry	6,962	17,803	28.11%
disgust	1,733	2,054	45.76%
fear	3,117	3,222	49.17%
happy	79,797	54,241	59.53%
neutral	33,708	40,858	45.21%
sad	11,175	14,156	44.12%
surprise	6,439	7,588	45.90%
Total	142,931	139,922	50.53%

Balanceado

Gender	Female	Male	% F
angry	6,962	6,962	50%
disgust	1,733	1,733	50%
fear	3,117	3,117	50%
happy	54,241	54,241	50%
neutral	33,708	33,708	50%
sad	11,175	11,175	50%
surprise	6,439	6,439	50%
Total	117,375	117,375	50%

Sesgo representacional

Gender	Female	Male	% F
angry	1,392	5,570	20%
disgust	347	1,386	20%
fear	623	2,494	20%
happy	10,848	43,393	20%
neutral	6,742	26,966	20%
sad	2,235	8,940	20%
surprise	1,288	5,151	20%
Total	23,475	93,900	20%

Sesgo estereotípico

Gender	Female	Male	% F
angry	1,392	5,570	20%
disgust	866	866	50%
fear	1,558	1,558	50%
happy	27,120	27,120	50%
neutral	16,854	16,854	50%
sad	5,587	5,587	50%
surprise	3,219	3,219	50%
Total	55,596	60,774	47.7%

Resultados de propagación

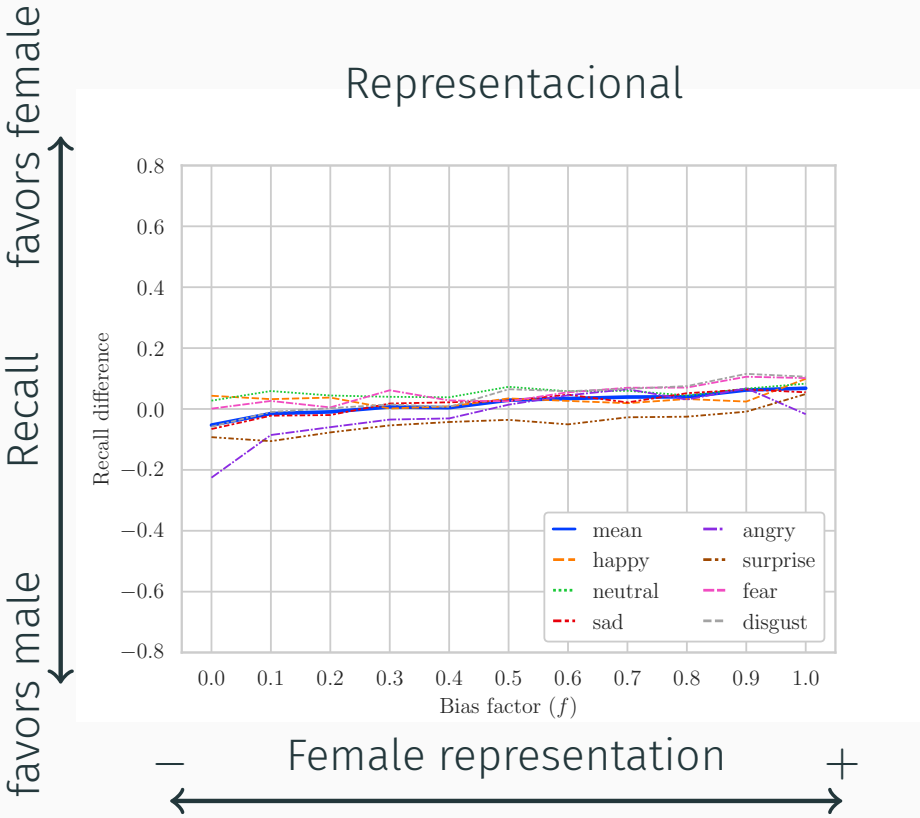


Figure 10: Diferencia en recall (recall en mujeres menos en hombres).

Resultados de propagación

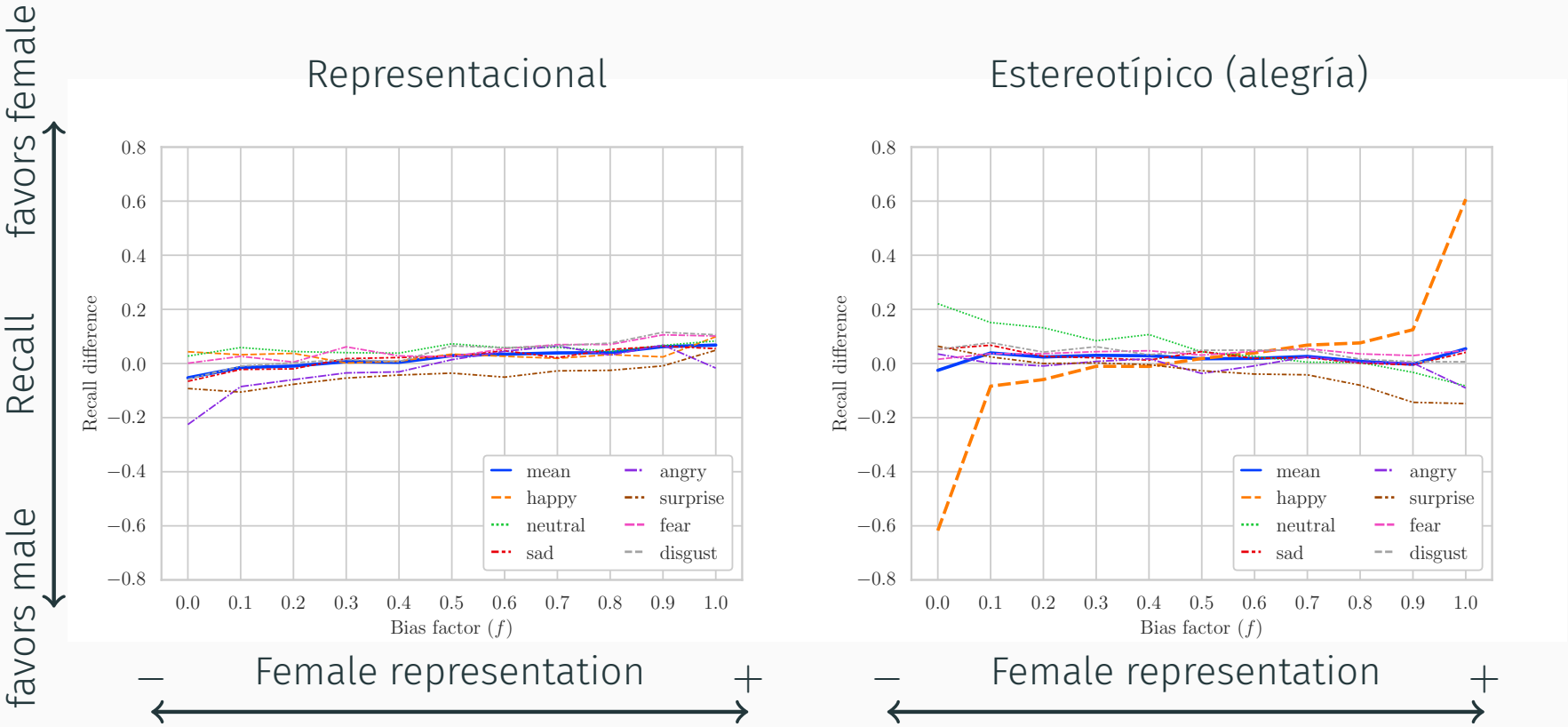


Figure 10: Diferencia en recall (recall en mujeres menos en hombres).

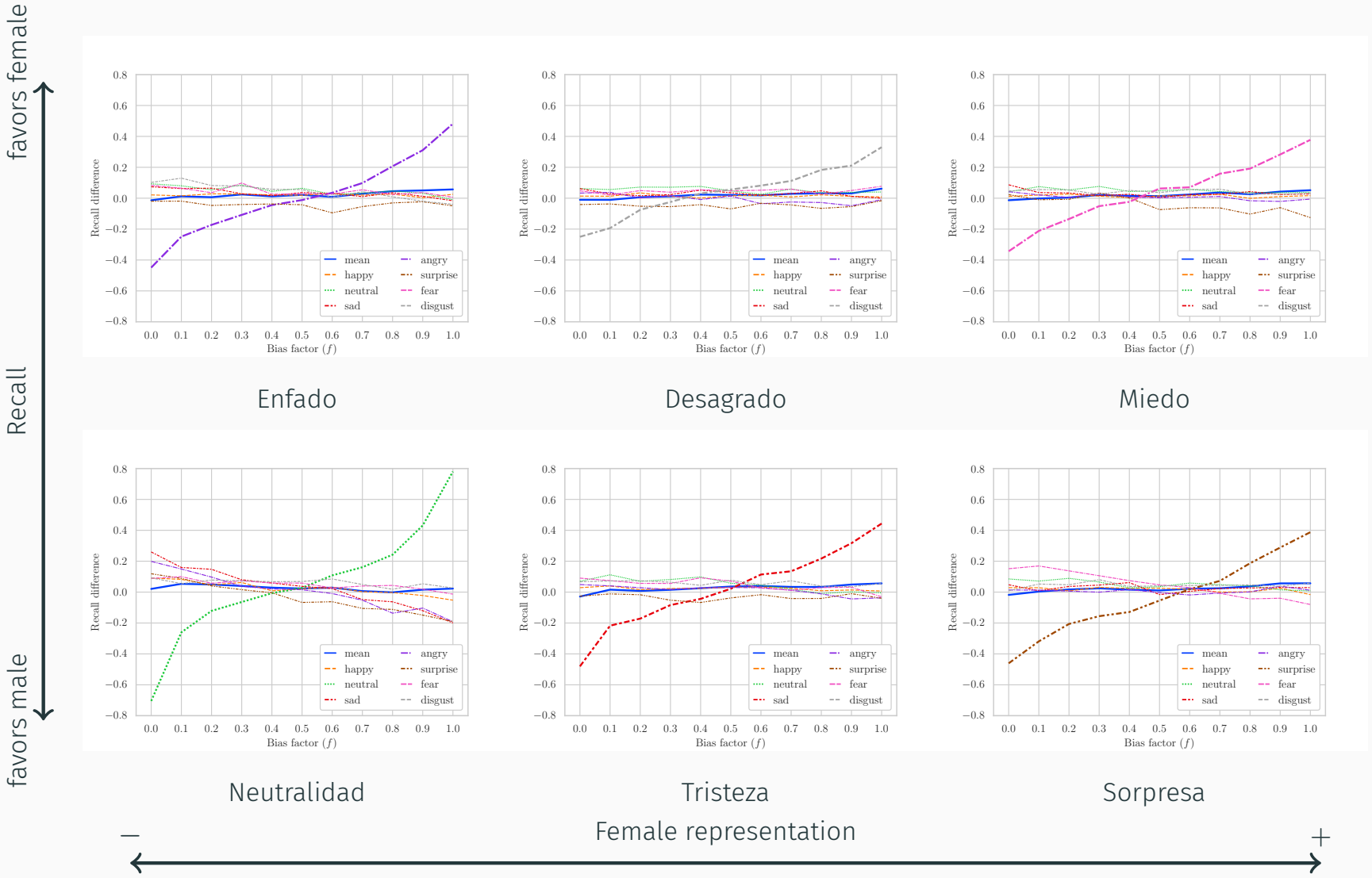


Figure 11: Efecto del sesgo estereotípico.

Propuestas

Measuring Transference from Dataset Bias to Model Predictions

- Extender el resultado anterior a **multigrupo** (raza).
- Desarrollar las métricas de sesgo en modelo necesarias.

- Extender el resultado anterior a **multigrupo** (raza).
- Desarrollar las métricas de sesgo en modelo necesarias.

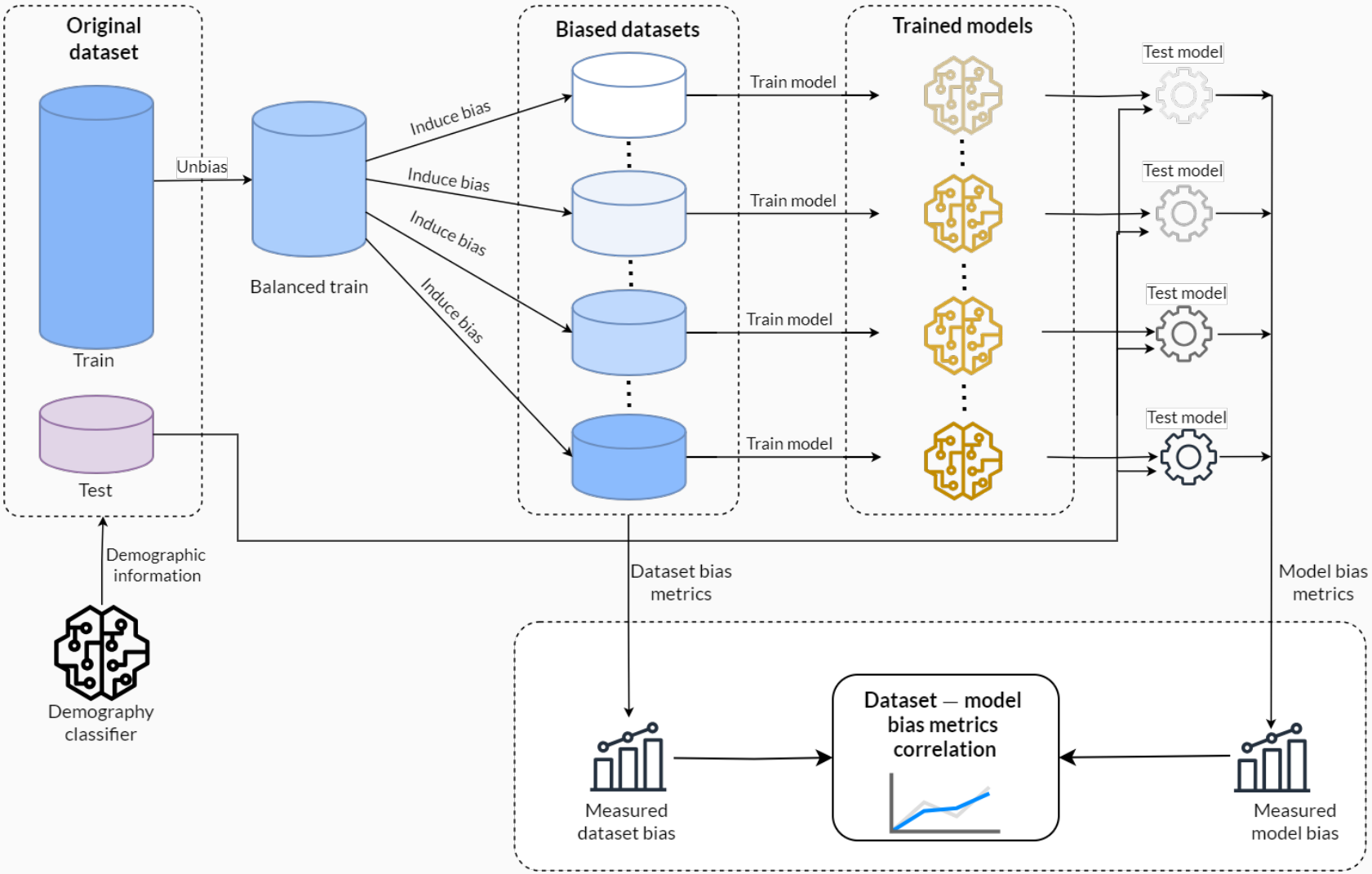


Figure 12: Resumen de la metodología

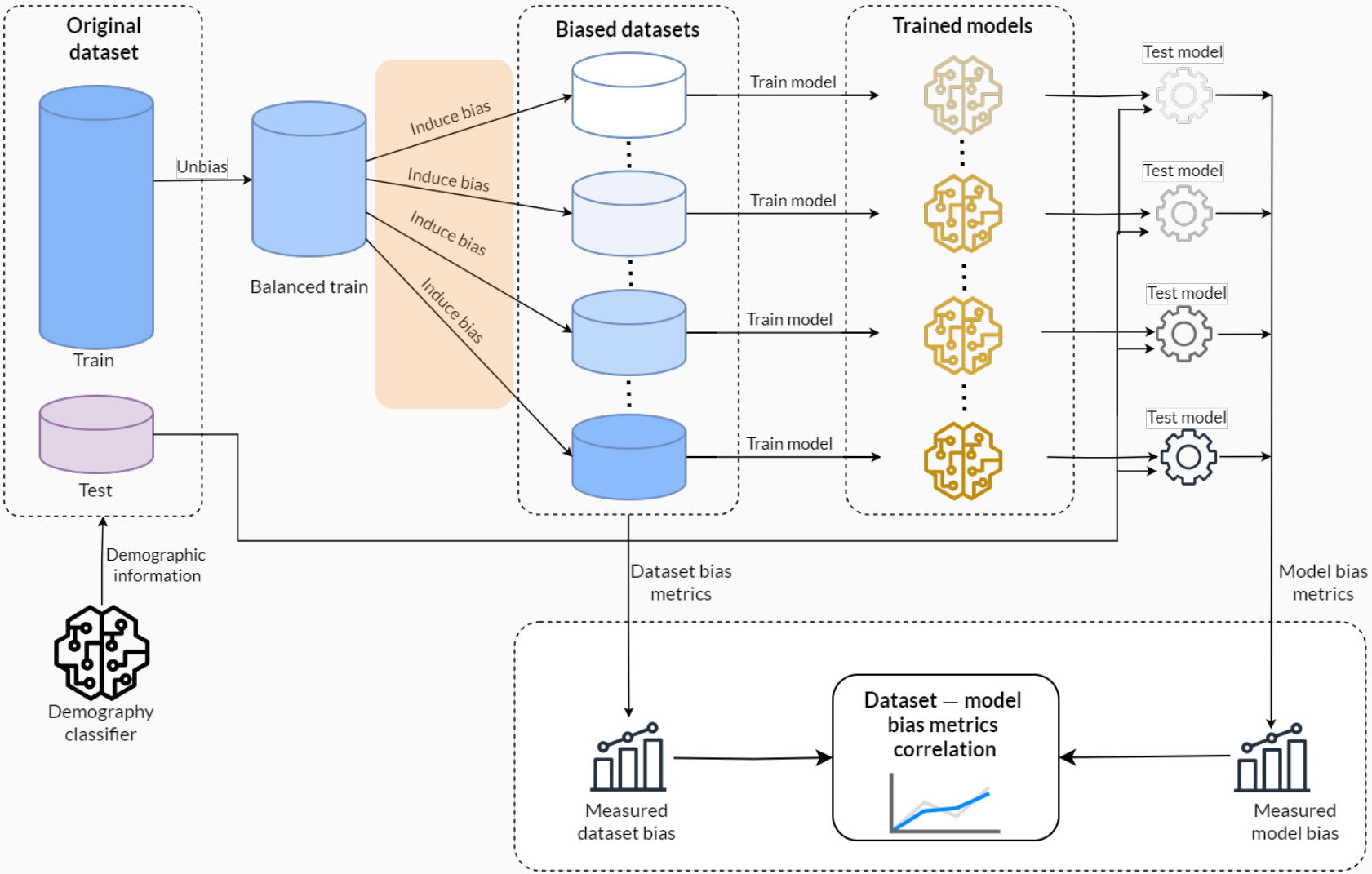


Figure 12: Resumen de la metodología

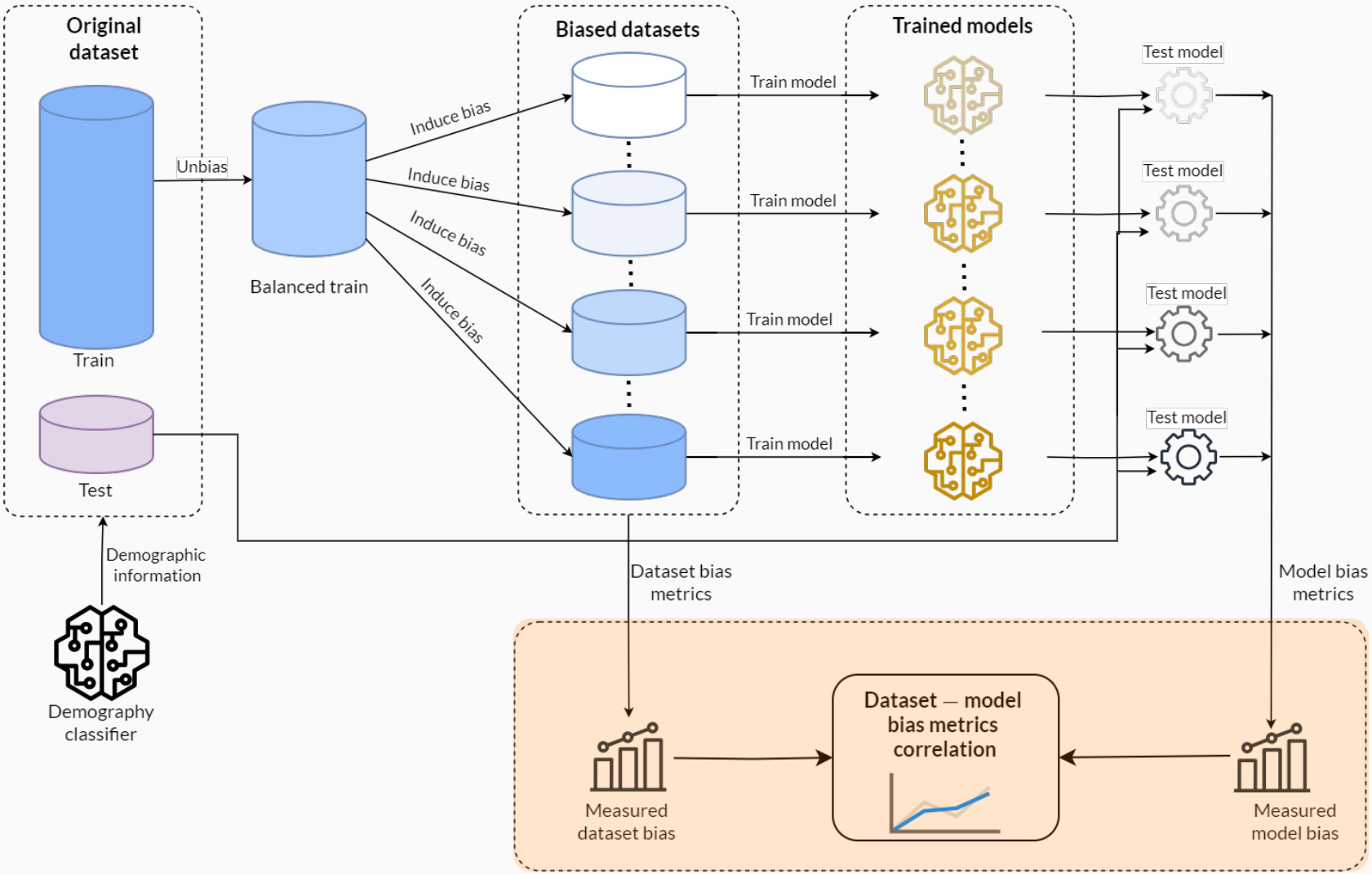


Figure 12: Resumen de la metodología

Métricas de sesgo en modelo

- Adaptaciones de[16]:
  - Term-by-term Multiclass Equalized Odds (**TTEqOdds**)
  - Classwise Multiclass Equalized Odds (**CEqOdds**)
  - Multiclass Equality of Opportunity (**EqOpp**)
  - Multiclass Demographic Parity (**DemPar**)
- Métricas previas:
  - Overall disparity (**OD**) [17]
  - Combined Error Variance (**CVE**) [18]
  - Symmetric Distance Error (**SDE**) [18]

---

[16] Preston Putzel and Scott Lee. *Blackbox Post-Processing for Multiclass Fairness*. 2022. arXiv: 2201.04461 [cs].

[18] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. "Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition". In: *Proceedings of the Workshop on Artificial Intelligence Safety 2022 (AISafety 2022)*. Vienna, Austria, 2022

[18] Cody Blakeney et al. "Measuring Bias and Fairness in Multiclass Classification". In: *2022 IEEE International Conference on Networking, Architecture and Storage (NAS)*. 2022, pp. 1–6

Metric correlation I

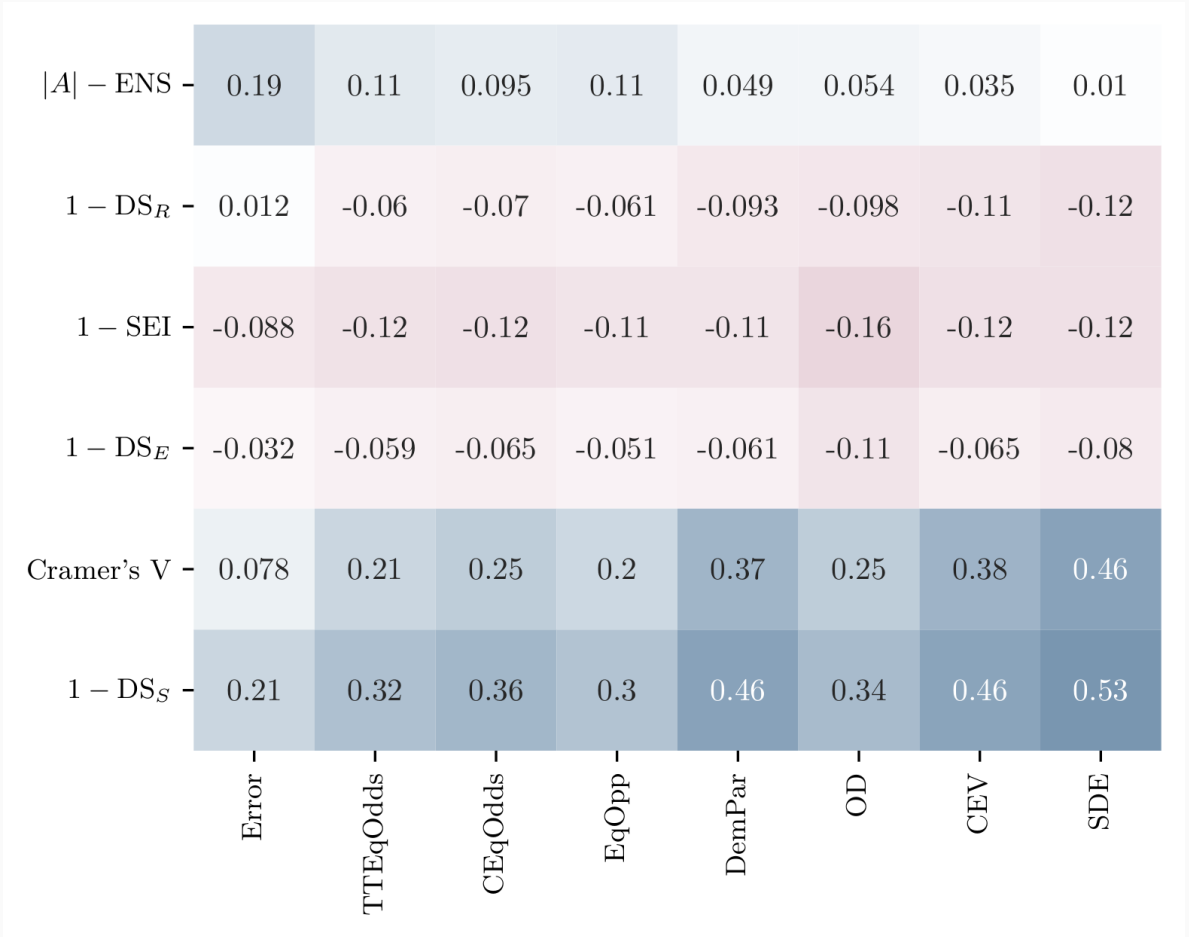


Figure 13: Spearman's ρ rank correlation between bias metrics.

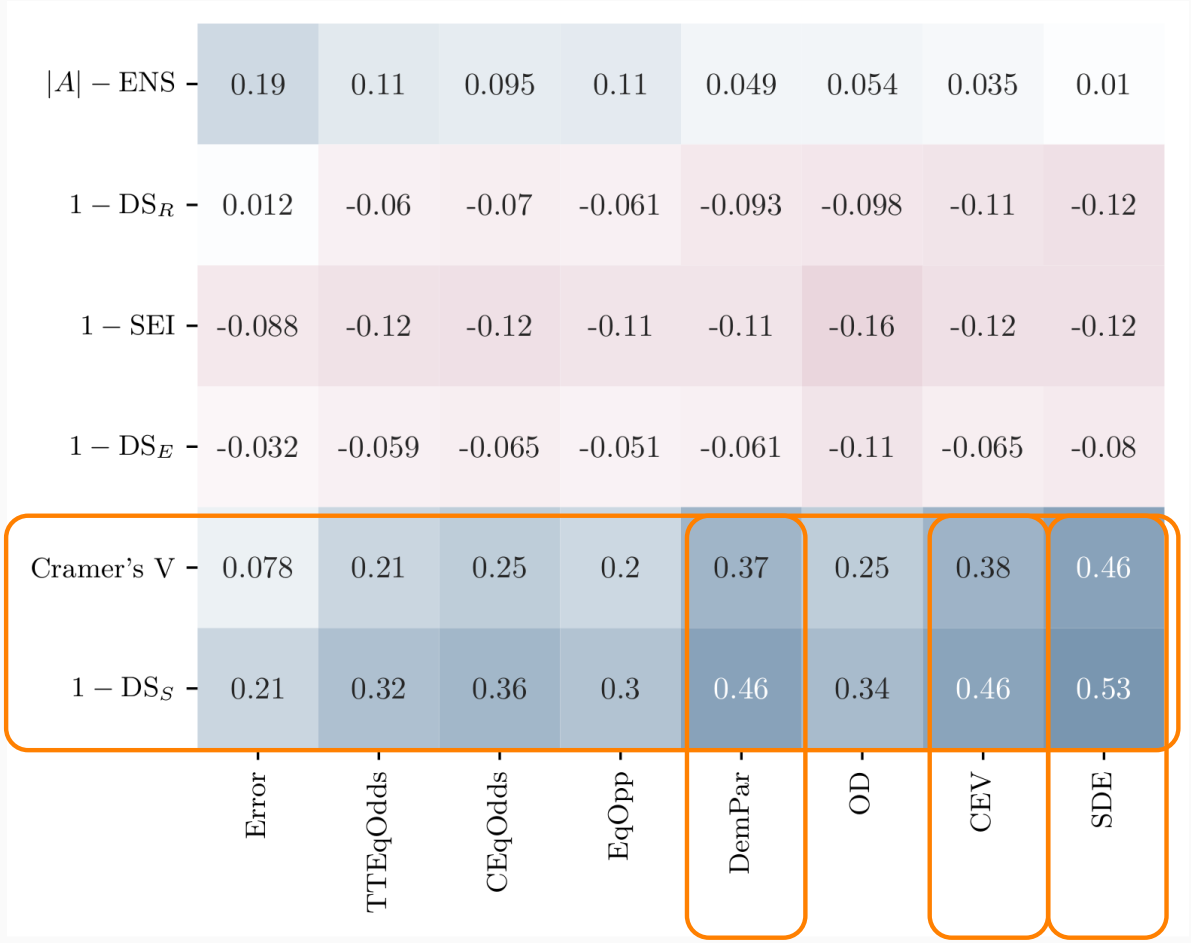


Figure 13: Spearman's ρ rank correlation between bias metrics.

Metric correlation II

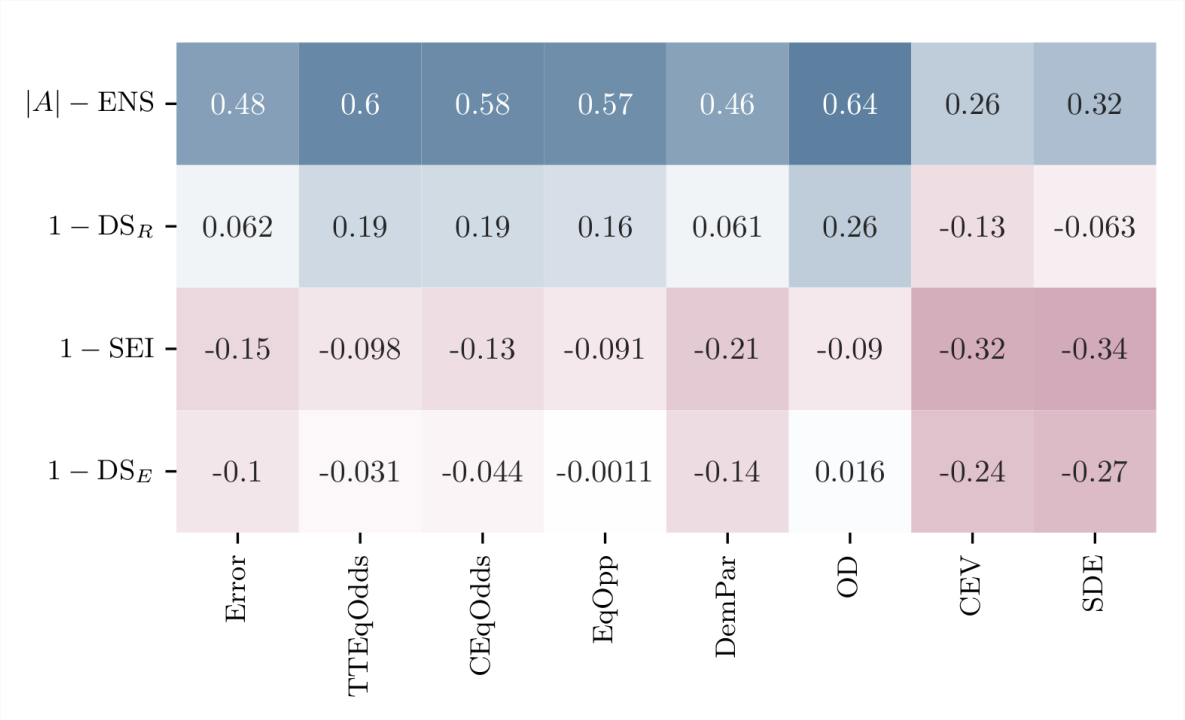


Figure 14: Spearman's ρ rank correlation between bias metrics, restricted to representational bias.

Metric correlation II

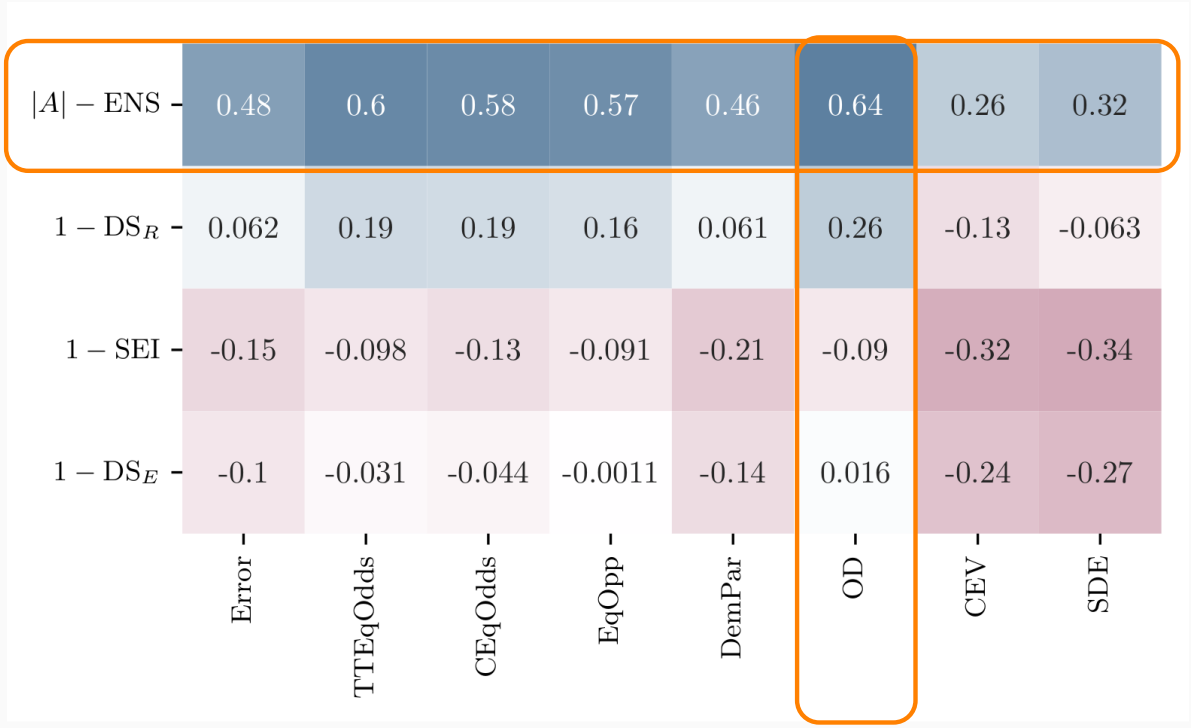


Figure 14: Spearman's ρ rank correlation between bias metrics, restricted to representational bias.

Introducción

Objetivos

Propuestas

Conclusiones y trabajo futuro

Conclusiones

En esta tesis:

- Hemos desarrollado una **taxonomía de tipos de sesgo en datasets** y métricas específicas para medirlo. También hemos desarrollado nuevas métricas de sesgo en modelos.
- Hemos diseñado la **metodología DSAP**, que permite comparar datasets y sirve como base para métricas más interpretables.
- Los datasets *In-The-Wild* (ITW) muestran un cambio hacia el **sesgo estereotípico**, que tiene un impacto más fuerte en los modelos que el sesgo representacional.
- Nuestros hallazgos subrayan la necesidad de **mitigar proactivamente los sesgos** en todas las etapas del aprendizaje automático.

¿Dónde estamos ahora?

- Estudio de sesgos en LAION.
- Mejora de los modelos auxiliares a través de *ensembles*.
- Sesgos en LLMs.

Gracias por la atención.

¿Preguntas?

✉ iris.dominguez@unavarra.es

https://irisai.neocities.org