Équité dans les recommandations d'emploi : estimer, expliquer et réduire les écarts entre les hommes et les femmes

G. Bied, P. Caillou, B. Crépon, C. Gaillac, M. Hoffmann, S. Nathan, E. Pérennes, M. Sebag

DARES, 12 Septembre 2023

- La présentation représente les opinions de ses auteurs. Elle ne représente pas nécessairement les points de vue ou les opinions de Pôle Emploi.
- L'algorithme audité ici n'est pas actuellement utilisé par Pôle Emploi.

Motivation

Apprenant d'un très large volume de données passées, les systèmes de recommandation peuvent améliorer les appariements sur le marché du travail en :

- réduisant le coût d'accès à l'information,
- suggérant des opportunités qui ne seraient pas considérées.

Ils présentent aussi des risques pour l'équité :

- les données mélangent les différents choix des demandeurs et recruteurs qui sont potentiellement genrés;
- les algorithmes optimisent un **objectif** qui n'est pas nécessairement celui des individus,

et cela peut amplifier les inégalités !

Contexte

Un partenariat depuis 2018 entre:

- CREST (GENES)
  B. Crépon, C. Gaillac, M. Hoffmann, E. Pérennès
- LISN (Université Paris Saclay), INRIA TAU Team
  P. Caillou, M. Sebag, G. Bied, S. Nathan
- Pôle emploi (DSEE)
  A. Bonnet, P. Beurnier, E. Chion, Y. De Coster, C. Nouveau, S. Robidou, C. Vessereau

Objectifs: *Valorisation des Données de Recherche d'Emploi*

1. Construire des systèmes de recommandation pour le marché de l'emploi dans le but de réduire le chômage frictionnel
2. Évaluer l'impact de ces systèmes
3. **Évaluer l'équité et les biais de ces systèmes** : documenter si des emplois différents sont recommandés aux femmes et aux hommes

Cette présentation et notre contribution

Le contexte expérimental et l'algorithme

Mesurer l'effet du genre sur les recommandations

- Le choix des variables d'intérêt
- Des premiers écarts "naïfs"

Mesurer l'effet du genre en contrôlant des préférences

- Une notion d'équité plus adaptée à notre contexte
- La méthodologie
- Des écarts "à profil professionnel et préférences égales"

Comparer les inégalités créées à celles pré-existantes

- Justification
- Résultats

Limiter ces inégalités à l'aide de méthodes adversariales

- Méthode
- Résultats et illustration du compromis

Cette présentation

Cette présentation suit:

- *Fairness in job recommendations: estimating, explaining, and reducing gender gaps* (G. Bied, C. Gaillac, M. Hoffmann, P. Caillou, B. Crépon, S. Nathan, M. Sebag), ECAI-workshop AEQUITAS 2023.

Peu de détails sur l'algorithme ici, mais tout est détaillé dans:

- *Toward Job Recommendation for All* (G. Bied, S. Nathan, E. Perennes, M. Hoffmann, P. Caillou, B. Crépon, C. Gaillac, M. Sebag), IJCAI 2023.
- *Designing Labor Market Recommender Systems: the Importance of Job Seeker Preferences and Competition* (G. Bied, P. Caillou, B. Crépon, C. Gaillac, E. Perennes, M. Sebag)

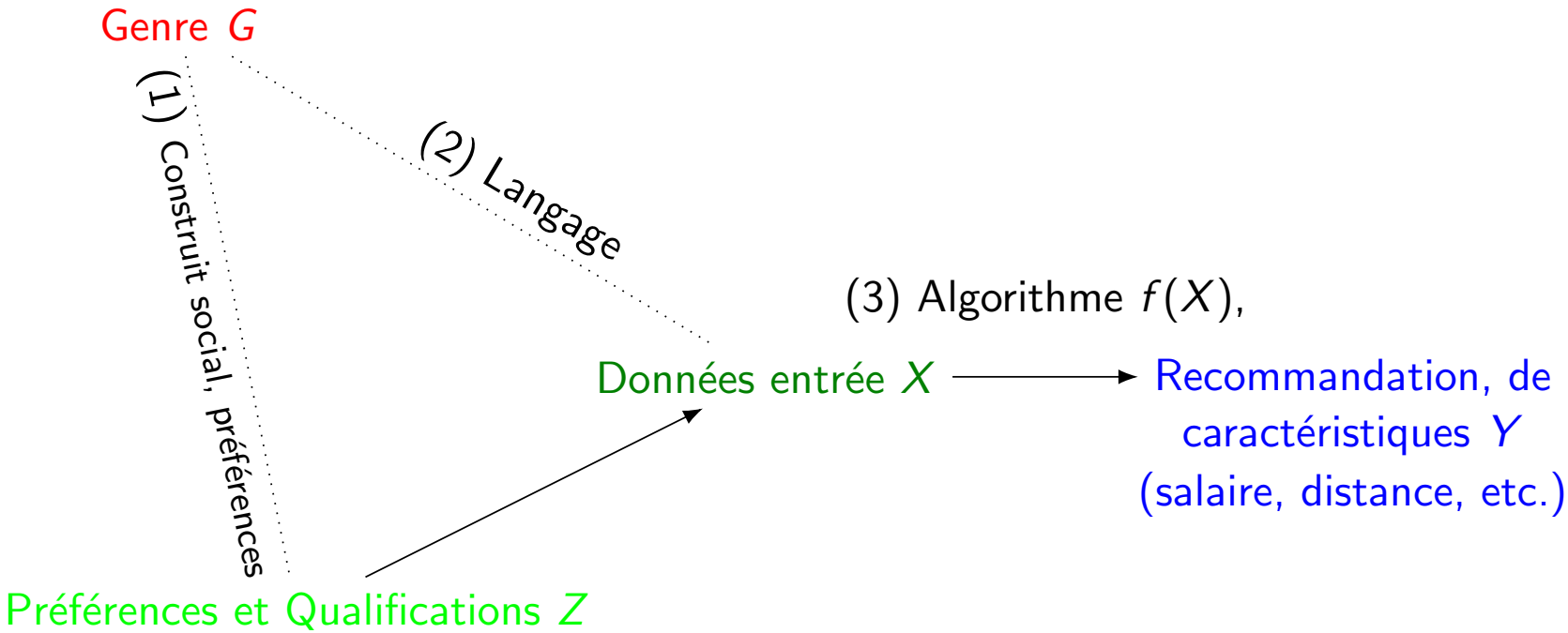
Le contexte expérimental et l'algorithme

- Région Auvergne-Rhône-Alpes, 2019-mi 2022
- 1.2M de demandeurs d'emploi, 2.2M offres, **285k embauches**
- Caractéristiques des DE et des offres, représentées comme des vecteurs de dimension ~ 500
- Incluent : profils, préférences, historique, texte.. vs salaire et conditions proposées, qualifications demandées, texte.
- *Le genre (binaire) est disponible mais ne constitue pas une donnée d'entrée de l'algorithme*
- Échantillon découpé en 85% / 15% test/train par semaine
- On se focalise sur tout les DE d'une certaine semaine du test (n=358,682) ainsi que tout les DE embauchés dans le test (n=41,787)

L'algorithme : principe

- Utilise les caractéristiques pour prédire les embauches et apprendre une notion de proximité entre les offres et les DE
- But: classer les offres selon leur chances de conduire à une embauche pour un demandeur (i.e. qu'il candidate et qu'il soit pris)
- Un réseau de neurones à deux niveaux :
  - Premier niveau de plongements sélectionne 1 000 offres d'emploi pour chaque DE
  - Second niveau qui reclasse les offres en utilisant un modèle plus fin et d'autres variables

Le problème



- La variable "sensible" est le genre G
- Les données d'entrée sont notées X...
- .. elles incluent les préférences (salaire demandé, etc.) et le profil professionnel Z
- Les caractéristiques des offres recommandées sont notées Y

Le choix des variables d'intérêt

La notion d'équité dépend de la **variables d'intérêt** Y, i.e du sens auquel on considère qu'il ne devrait pas y avoir d'inégalités entre les recommandations faites aux femmes et aux hommes.

Le choix des variables d'intérêt

La notion d'équité dépend de la **variables d'intérêt** Y, i.e du sens auquel on considère qu'il ne devrait pas y avoir d'inégalités entre les recommandations faites aux femmes et aux hommes.

- Une mesure de la performance dans le sens de **l'objectif de l'algo** est le recall@k : la proportion des embauches qui sont correctement classées dans le top k recommandé.
  $\Longrightarrow$ Objectif mesurable avec nos données
- Entre dans l'objectif **des demandeurs**, mais un concept plus général serait **l'(espérance d')utilité** qu'ils tirent des recommandations
  - Objectif qui n'est pas directement mesurable
  - **Un proxy** est de comparer les différentes caractéristiques des offres (salaire, distance, qualification, type de contrat...)
  - et une mesure de **l'adéquation** aux critères de recherche

Des premiers écarts "naïfs"

- On regarde les différences entre les femmes et les hommes

$$\delta = \mathbb{E}[Y|G = \text{Femme}] - \mathbb{E}[Y|G = \text{Homme}].$$

- La performance des recos est meilleure pour les femmes :
  - l'algo. classe correctement dans ses 20 premières recos l'offre d'emploi sur laquelle un DE a été embauché dans 35% des cas, 33,3% pour les hommes et 36,6% pour les femmes.
- Les emplois recommandés sont statistiquement différents

	$\hat{\delta}$ (Entier)	$\hat{\delta}$ (Support commun)
Salaire (log)	-0.023***	-0.016***
Distance (km)	-0.474***	-0.231***
Cadre	-0.004***	-0.002***
CDI	-0.040***	-0.034***
%Femme < 20	-0.411***	-0.219***
Heures travaillées/s	-2.934***	-1.957***
Adéquation aux param.	-0.028***	-0.019***

Notes: La colonne 1 montre $\hat{\delta}$ pour les DE d'une semaine (n=358,682); la 2 montre $\hat{\delta}$ pour le support commun, conservant les individus dont le score de propension est compris entre $[0,05; 0,95]$ (n=228,625).

Une notion d'équité plus adaptée à notre contexte

- La littérature économique documente des différences femmes/hommes de préférences pour le temps de trajet, la flexibilité du contrat, le salaire...
- Si l'utilité était observée, **une partie** de ce qui apparaît comme des inégalités sur les différentes caractéristiques des recos serait gommée
- Une notion moins contraignante d'équité : les recos genrées peuvent être considérées comme équitables, à condition que les écarts soient conformes aux préférences des individus.
- Reste que **l'autre partie** des différences vient bien d'une valorisation genrée (et pas équitable) de l'algorithme des différents profils.

La méthodologie

- Sous certaines conditions, notre écart moyen peut se décomposer (*Oaxaca*):
  1. en un effet expliqué par les profils pro & préférences Z;
  2. et un reste τ qui est ce résidu des différences de recos qui ne peut pas être expliqué par Z, et n'est pas "juste".
- Le modèle : $Y = \tau G + \mu_0(Z) + \varepsilon, \quad E(\varepsilon|Z, G) = 0.$

La méthodologie

- Sous certaines conditions, notre écart moyen peut se décomposer (*Oaxaca*):
  1. en un effet expliqué par les profils pro & préférences Z;
  2. et un reste τ qui est ce résidu des différences de recos qui ne peut pas être expliqué par Z, et n'est pas "juste".
- Le modèle : $Y = \tau G + \mu_0(Z) + \varepsilon, \quad E(\varepsilon|Z, G) = 0.$
- L'estimation du biais de genre τ est faite de manière flexible et en limitant l'apriori en utilisant le *Double Machine Learning*
- Z comprend : le salaire de réserve, le type de contrat, le secteur, le temps de trajet désirés ainsi que les qualifications, compétences, département et l'expérience.

Des écarts "à profil professionnel et préférences égales"

	$\hat{\delta}$	$\hat{\tau}$
Salaire (log)	-0.016***	-0.004***
Distance (km)	-0.231***	0.400***
Cadre	-0.009***	-0.002**
CDI	-0.034***	-0.014***
%Femme $<$ 20	-0.219***	-0.033***
Heures travaillées/s	-1.957***	-0.381***
Adéquation aux param.	-0.019***	-0.011***

Notes: les 1 et 2 montrent $\hat{\delta}$ et $\hat{\tau}$ pour le support commun (n=228,625).

- Les différences sont réduites mais subsistent
- 25 % de l'écart salarial reste inexpliqué par le profil et les préférences

Comparer les inégalités créées à celles pré-existantes

- Même si l'on contrôle pour les préférences, ce n'est toujours qu'un proxy pour l'utilité des DE.. qui peut expliquer ces différences
- On peut lier sous certaines conditions les choix de candidatures à l'utilité que les demandeurs espèrent tirer d'une potentielle embauche sur une offre considérée Modèle
- Une comparaison naturelle est donc les inégalités observées entre les offres sur lesquelles **les DE candidatent** τ_{Cand}.
- On calcule l'écart entre les inégalités de l'algorithme et celles des candidatures $\tau_{\text{DifC}} = \tau - \tau_{\text{Cand}}$ (négatif = on n'augmente pas ces inégalités)
- Aussi avec les inégalités dans les embauches (cf. papier)

Comparer les inégalités crées à celles pré-existantes

Candidatures	**Différences Femmes/Hommes**				**Différence de Différences**	
	τ_{Cand} (Obs.)	p-value	τ	p-value	τ_{DifC}	p-value
Salaire (log)	-0.012	0.000	-0.011	0.000	0.002	0.559
Distance (km)	-4.338	0.000	0.524	0.002	4.905	0.000
Cadre	-0.002	0.322	-0.002	0.607	0.001	0.791
CDI	-0.023	0.003	-0.021	0.052	0.002	0.900
%Femmes < 20	-0.142	0.000	-0.067	0.000	0.076	0.000
Heures travaillées/s	-1.177	0.000	-0.675	0.000	0.507	0.001
Adéquation aux param.	-0.029	0.000	-0.025	0.000	0.007	0.156

Notes: Les résultats sont présentés sur le sous-échantillon des demandeurs d'emploi embauchés pour pouvoir être comparées avec les caractéristiques des embauches. En raison de sources de données différentes, nous étudions la sous-population des demandeurs d'emploi embauchés au cours des semaines de test pour lesquelles nous observons des candidatures (toutes semaines confondues). La première colonne présente les estimations conditionnelles des écarts entre les femmes et les hommes sur les candidatures observée pour la population sur le support commun. La troisième colonne présente la même différence sur les caractéristiques des recommandations de l'algorithme. La cinquième colonne rapporte la différence de ces deux dernières différences, c'est-à-dire les estimations conditionnelles des différences entre les caractéristiques d'une candidature et la recommandation de l'algorithme.

- Les recos ne diffèrent pas de là où les DE candidatent (et sont embauchés), sauf pour le nombre d'heures de travail et le type de secteur, où l'algo réduit ces écarts.

Limiter ces inégalités à l'aide de méthodes adversariales

- Même avec cette notion plus faible d'équité tenant compte des préférences, il subsiste des différences. Pour les limiter, nous étudions une solution "adversariale".
- Le principe est d'entraîner un modèle "adversaire", qui a pour objectif L_{adv} de prédire le genre à partir des recommandations qui sont faites.
- S'il y arrive, l'algorithme initial $L_{classif}$ voit son objectif initial pénalisé. Il intègre donc dans son objectif le fait qu'il ne faut pas que l'adversaire réussisse à deviner le genre à partir des recommandations qu'il émet.
- Ce compromis entre l'objectif initial et celui de l'adversaire est indexé par un paramètre λ. On utilise : $(L_{classif} - \lambda L_{adv})$.
  - Pour $\lambda = 0$, l'algorithme maximise les chances d'observer des embauches sur les recos
  - Pour λ grand, l'algorithme accepte de diminuer ces chances pour ne pas créer de différences de genre

Résultats et illustration du compromis

	$\lambda = 0$	p-value	$\lambda = 0.01$	p-value	$\lambda = 1$	p-value
Performance						
Recall@20	0.351		0.346		0.335	
Recall@20 (hommes)	0.333		0.329		0.320	
Recall@20 (femmes)	0.366		0.361		0.348	
Exactitude de l'adversaire			0.784		0.530	
Écarts condi. Z						
Salaire (log)	-0.005	0.014	-0.001	0.035	-0.001	0.110
Distance	0.542	0.000	0.059	0.016	0.100	0.000
Cadre	-0.002	0.319	-0.001	0.177	-0.001	0.052
CDI	-0.027	0.001	-0.005	0.001	-0.006	0.000
%Femme < 20	-0.058	0.000	-0.009	0.000	-0.012	0.000
Heures travail.	-0.695	0.000	-0.103	0.000	-0.132	0.000
Adéq. param.	-0.022	0.000	-0.003	0.000	-0.003	0.000

Notes : Les résultats sont présentés sur le sous-échantillon des demandeurs d'emploi embauchés, pour différents poids λ donnés au terme adversarial dans la fonction de perte. Les indicateurs de performance sont calculés sur l'ensemble de la population. Les écarts inconditionnels correspondent à la différence de moyenne entre les hommes et les femmes sur le sous-échantillon des demandeurs d'emploi embauchés. Les écarts conditionnels sont calculés sur la population des demandeurs d'emploi embauchés ayant des caractéristiques suffisamment comparables.

- La performance est réduite à mesure que λ augmente
- Les écarts conditionnels (et inconditionnels) sont réduits sans être supprimés

Conclusion et perspectives

- Développe une méthodologie pour l'analyse des inégalités de genre potentiellement générées par l'algorithme
- Les recos ont une meilleure performance pour les femmes, mais à paramètres de recherche égaux, les femmes se voient recommander des offres qui satisfont leur critères de recherche moins souvent.
- Différences qui ne sont pas plus prononcées que celles que l'on observe déjà dans les candidatures et les embauches, suggérant que l'algorithme n'augmente pas, voire réduit, les biais genrés par rapport aux comportements !

Conclusion et perspectives

- Nous proposons une solution pour gommer ces différences, qui a toutefois un coût en termes de performances pour toute la population, et en particulier pour les femmes.
- Cette analyse illustre la question de l'importance de l'objectif utilisé et du référentiel selon lequel on juge de l'équité des recommandations
- Les performances d'un algorithme qui tenterait de "débiaiser" les recommandations se heurtent à la question de leur acceptabilité par les individus, si cela va à l'encontre de là où ils auraient spontanément candidaté

Machine learning algorithm - job seeker features (in Z)

Preferences	
Reservation wage (euros / hour)	numeric
The job seeker is looking for a full-time job	binary
Target job sector	categorical (x14)
Target job	categorical (x110)
Target type of contract	categorical (x13)
Maximum commuting time	numeric
Maximum (and Minimum) number of work hours per week	numeric
Qualifications	
Number of years of experience	numeric
Maximum level of qualification	categorical (x10)
Department	categorical (x13)
Vocational training field	categorical (x27)
Skills (SVD)	numeric (x50)
Driving licences	categorical (x22)
Number of languages spoken	numeric
Means of transportation	categorical (x5)

Machine learning algorithm - job seeker features (not in Z)

Socio-demographic variables	
Number of children	numeric
Jobseeker lives in a QPV area	numeric
Past employment history	
Number of unemployment periods in lifetime	numeric
Reason why the job seeker registered at PES	categorical (x15)
Type of accompaniment received from PES	categorical (x4)
Main obstacles assumed to slow return to employment	categorical (x4)
Resume	
Curriculum text (SVD)	numeric (x100)
Number of words in the curriculum text	numeric
Number of visit cards	numeric
Number of sectors considered by the job seeker	numeric
Geographic information	
Firm density within zip code	numeric
Unemployment rate within zip code	numeric
Latitude	numeric
Longitude	numeric

The Overlap Assumption

- Impact τ needs be assessed from men and women "sufficiently close" due to the *common support* assumption.
- In practice:
  - Fit random forest to predict gender based on main search & socio-demographic characteristics (88% accuracy);
  - Common support defined as predictions between 0.05 and 0.95.

A simple model of application behavior

Back

- Job seeker x, job ad z
- Chance of being hired if they apply: $p(x, z)$
- $\pi(x, z)$: job seekers' estimate of $p(x, z)$
- Application cost k
- If they apply: utility $U(x, z) + \varepsilon - k$ if hired, $U_0(x) - k$ if not;
- Else: $U_0(x)$

A simple model of application behavior

Back

- Job seeker x, job ad z
- Chance of being hired if they apply: $p(x, z)$
- $\pi(x, z)$: job seekers' estimate of $p(x, z)$
- Application cost k
- If they apply: utility $U(x, z) + \varepsilon - k$ if hired, $U_0(x) - k$ if not;
- Else: $U_0(x)$
- Decision of application:

$$\underbrace{\pi(x,z)(U(x,z)+\epsilon)+(1-\pi(x,z))U_0(x)-k}_{\text{Expected utility when applying}} \geq \underbrace{U_0(x)}_{\text{Utility without applying}} .$$

- Probability of application:

$$A(x,z) = F_{-\epsilon}\left(U(x,z) - U_0(x) - \frac{k}{\pi(x,z)}\right)$$

- Probability of hire $H(x, z) = A(x, z)p(x, z)$
  - In the paper: elaborate model of a two-sided market with transferable utility and an application stage