Utilisation des données mobiles pour prédire des indicateurs économiques

Ayizou R.[1], Clochard G.[2], Diallo H.[1],

Fall O. [3], Hollard G.[4], Sene O.[5]

[1]ENSAE de Dakar, [2]Université de Chicago, [3]ANSD,

[4]Crest, Ecole Polytechnique et CNRS, [5]Université Alioune DIOP de Bambey

2 Décembre 2022

PLAN

PLAN

Le projet GUISSTANN

GUIder la Statistique publique Sénégalaise, Téléphonie, Algorithmique et Nouvelles techniques Numériques

- Liste des partenaires

ANSD
Agence Nationale de la Statistique et de la Démographie

ÉCOLE POLYTECHNIQUE
IP PARIS

sonatel
orange

UADB

Description sommaire du projet GUISSTANN

- Utiliser des données téléphonique et mobile money pour prédire des indicateurs issus du recensement de la population
- Utilisation d'algorithmes de machine learning pour la prédiction
- Intérêt scientifique : utiliser une source de données abondante et bon marché pour améliorer la statistique publique au Sénégal

PLAN

Sources alternatives de données

- Les données "alternatives" (CDR, mobile money, satellite, etc) permettent de disposer massives et de bonne qualité
- Plusieurs exemples montrent que ces données permettent de faire des prédictions de bonne qualité, rapidement et à moindre coût
- Le principe élémentaire est de disposer de données individuelles "au sol" qui soient appareillées avec des données téléphoniques
- L'apprentissage statistique permet alors de prévoir les indicateurs "au sol" à partir des données "telco"

Machine learning et développement

Les données alternatives ont été utilisées avec succès dans les pays en développement, notamment pour:

- Etablir des cartes de pauvreté (Blumenstock et al. 2015)
- Améliorer le ciblage des populations en difficulté (Aiken et al. 2022)
- Etudier la mobilité (Erfani and Frias-Martinez, 2022)

Quelle est la contribution de GUISSTANN ?

- Beaucoup plus d'indicateurs disponibles (santé, éducation, mobilité de long terme, production agricole, etc)
- Possibilité d'évaluer la qualité de la prédiction à partir du recensement (exhaustif!) de 2023
- Produire une liste des indicateurs qui peuvent être prédits ou non par les CDR (ex: richesse, éducation, culture du bissap)

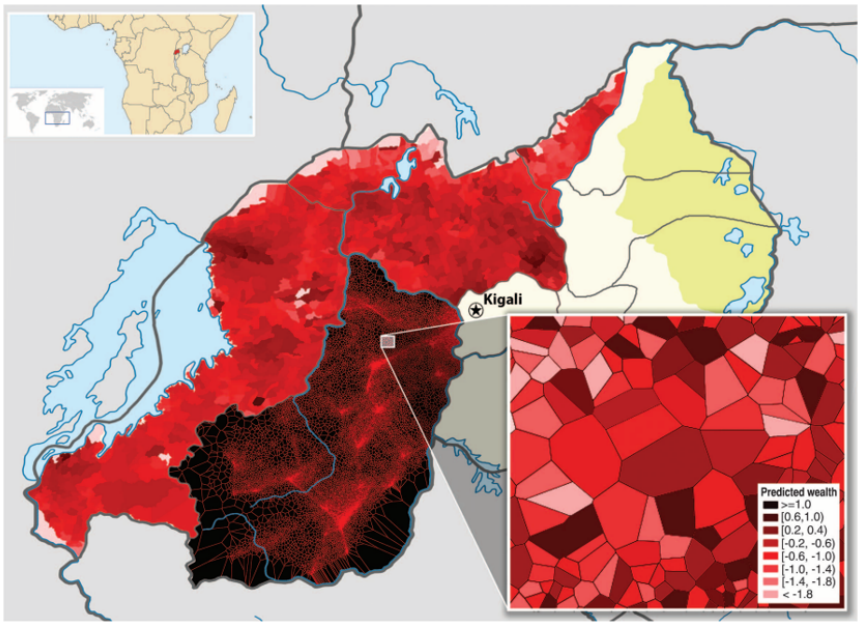
Intérêt pour la statistique sénégalaise

Couplage données téléphonique/données terrain:

- Mises-à-jour plus fréquentes des indicateurs du recensement
- Guider les enquêtes de terrain en fonction de la qualité des prédictions
- Créer une culture de la *qualité* des données de terrain (plus que de la quantité)

Exemple du Rwanda (Blumenstock et al. 2015)

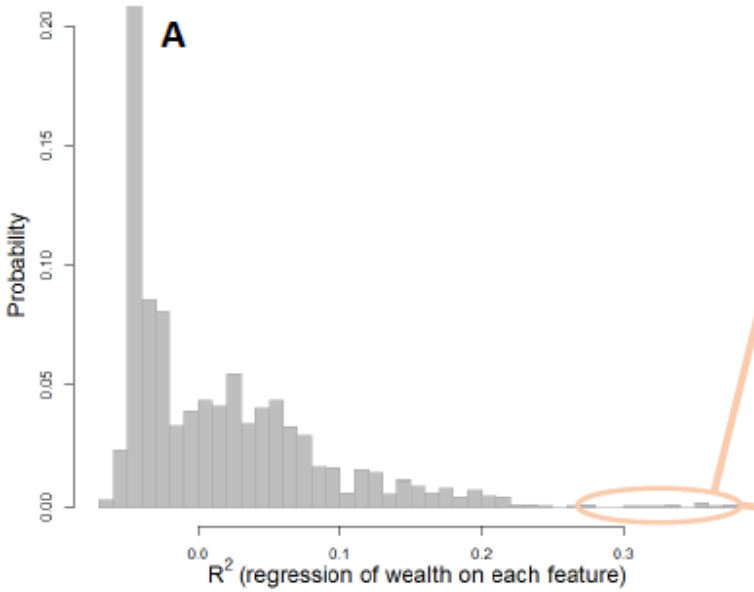
- La carte ci-dessous est obtenue en utilisant du ML à partir d'une enquête téléphonique auprès de ≤ 1000 individus
- La qualité de la prédiction à l'échelle du pays est estimée à partir des données DHS



Méthodologie

- Utilisation d'un automate (DFA) pour construire environ 2500 variables à partir des CDR
- Apprentissage à partir du petit échantillon (n=856) pour lequel on dispose simultanément des données CDR et de pauvreté
- Identification d'un modèle avec un nombre raisonnable de variables (une centaine) pour limiter le sur-apprentissage
- Utilisation du modèle pour extrapoler à l'ensemble du pays
- Évaluation de la qualité de la prévision à partir des données disponibles (DHS dans le cas décrit)

Variables clés identifiées par le modèle



B	Feature	R^2
	Weighted average of all first-degree neighbor's "Day of week (DoW) entropy" of outgoing SMS volume	0.376
	Avg. of alter's "Hour of Day (HoD) entropy", outgoing SMS	0.371
	Avg. of alter's DoW entropy for incoming SMS	0.364
	Avg. of alter's HoD entropy for incoming SMS	0.356
	Avg. of alter's HoD entropy for incoming calls over 60s	0.354
	Avg. of alter's DoW entropy for incoming calls over 60s	0.351
	Avg. of alter's HoD entropy for outgoing calls over 60s	0.338
	Avg. of alter's DoW entropy for outgoing calls over 60s	0.330
	Avg. of alter's number of unique call contacts	0.330
	Avg. of alter's number of unique SMS contacts	0.318

Figure: Variables sélectionnées dans Blumenstock et al. (2015)

Valeur ajoutée du ML par rapport à l'intuition

- Une question importante est l'apport du ML comparé à l'intuition des chercheurs
  - Si les individus qui dépensent le plus en téléphonie sont aussi les plus riches, a-t-on vraiment appris quelque chose ?
- Blumentstock et al. (2015) trouvent que le meilleur "feature" a une meilleure capacité prédictive que les modèles utilisant les variables "intuitives"

A	Elastic Net		Random Forest	
	r	R^2	r	R^2
Optimal DFA-based model	0.68	0.46	0.63	0.40
"Intuitive" 5-feature model	0.44	0.20	0.37	0.14
Single-feature model	0.61	0.38	0.46	0.22

PLAN

Comment évaluer l'aspect éthique et assurer la sécurité des données?

L'appariement des données Sonatel et ANSD introduit un risque pour la confidentialité des données personnelles

- Une attention particulière est portée à la protection des données (architecture informatique dédiée et accès très restreint au Sénégal uniquement, pseudonymisation, etc)
- L'aspect éthique sera évalué par toutes les instances juridiques et académiques compétentes

Cadre légal au Sénégal

- La loi statistique explicite les principes fondamentaux de la statistique publique dont le Secret statistique et l'utilisation exclusive à des fins statistiques
- Ainsi que la loi portant sur la Protection des données à caractère personnel

Consentement des individus

- L'aval de la Commission de Protection des données à Caractère personnel -CDP
- L'article 20 de la **loi n° 2008-12 du 25 janvier 2008 portant sur la Protection des données à caractère personnel** que les traitements des données à caractère personnel ayant un motif d'intérêt public notamment à des fins historiques, statistiques ou scientifiques sont mis en œuvre après autorisation de la Commission des Données Personnelles.

Biais Algorithmiques

- Apprentissage à partir des individus ayant un abonnement actif à la Sonatel
- Mauvaise prise en compte des personnes n'ayant pas de téléphone

PLAN

Travail en cours:

- Validation de l'architecture conjointe entre ANSD et Sonatel
- Dossier pour la CDP
- Accès aux données ANSD par les étudiants, données échantillons de CDR
- Validation éthique par IRB

→ Ces étapes devraient être terminées pour fin décembre

PLAN

Conclusion

Perspectives scientifiques:

- Guider l'effort de collecte sur le terrain en identifiant les aspects les moins bien prédits (complémentarité entre enquête de terrain et prédiction à partir des CDR)
- Aller de la statistique descriptive vers l'évaluation d'impact via les CDR ?

Restitution prévue en mars avec ouverture à des nouveaux partenaires et de nouvelles problématiques