Approche dynamique de la ségrégation: une analyse à partir de données de téléphonie mobile

Lino Galiana (D2E)

En collaboration avec Benjamin Sakarovitch (SSP-lab), François Sémécurbe (SSP-lab) et Zbigniew Smoreda (Orange Labs)

Séminaire du Département des Etudes Economiques

19 février 2019

Introduction

Une photographie de la distribution du revenu (Marseille)

- Données fiscales: concentration spatiale des premiers et derniers déciles à Marseille
  - Quelle dynamique infra-journalière ?

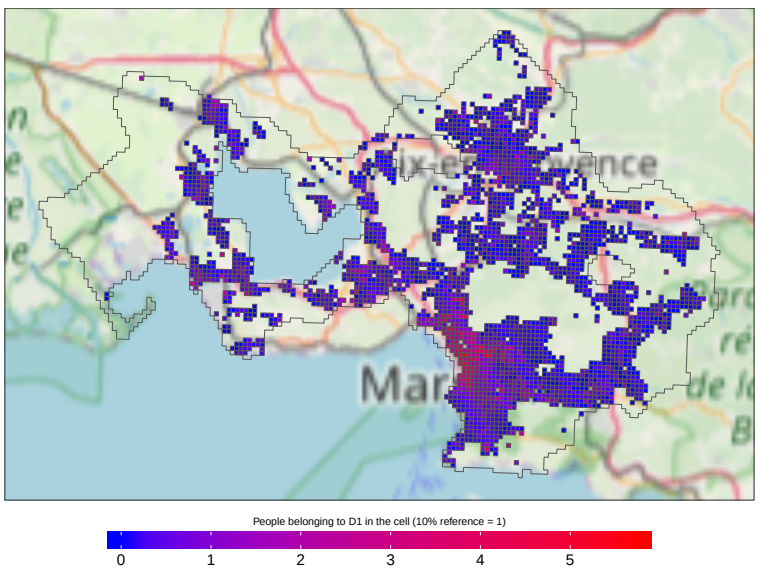


Figure 1: Répartition spatiale des ménages du premier décile

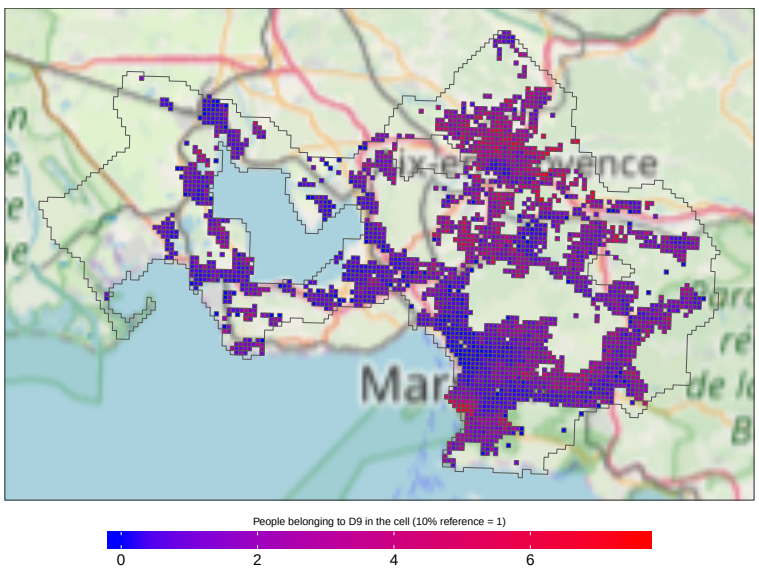


Figure 2: Répartition spatiale des ménages du dernier décile

Pourquoi adopter une vision dynamique de la ségrégation ?

- Ségrégation n'est pas un processus figé:
  - Dimension temporelle: composition d'un quartier évolue dans le temps
  - Résidence ne représente pas nécessairement l'endroit où l'individu passe du temps

- Questions de recherche:
  - Comment les mobilités affectent-elles la ségrégation urbaine ?
  - La ségrégation résidentielle masque-t-elle une dynamique infra-journalière?

- Dynamique infra-journalière marquée:
  - Davis et al. (2017): ségrégation lieux de sortie (restaurants) 50% inférieure ségrégation résidentielle
  - Le Roux et al. (2017): ségrégation jour inférieure de 15 à 30% à la ségrégation nocturne

Pourquoi utiliser des données de téléphonie pour étudier la ségrégation?

- Ségrégation principalement étudiée à partir d'indices
  - Indice de dissimilarité (Duncan & Duncan, 1955)

$$ID = \frac{1}{2} \sum_{j=1}^{J} \left| \frac{w_j}{W_T} - \frac{n_j - w_j}{N_T - W_T} \right|$$

- Données administratives $\Rightarrow$ ségrégation résidentielle:
  - Vision statique de la ségrégation
  - Séparation groupes sociaux dans l'espace résidentiel
  - Pas d'information sur lieux fréquentés

- Mobilités affectent de manière continue la distribution spatiale du revenu
  - Besoin de données géolocalisées à haute-fréquence...
  - ... qui doivent être combinées aux données classiques pour caractériser individus

Approche adoptée

- On se propose d'étudier la dynamique infra-journalière:
  - 48 points: 24 pour les jours de semaine, 24 pour le weekend
- Implique de construire des indices de ségrégation dépendant du temps
  - Construire une série d'indices de dissimilarité pour chaque ville
- Champs: unités urbaines (UU) de Lyon et Marseille
  - `Filosofi`: ménages dont domicile dans les limites UU
  - `Téléphonie`: ménages dont domicile simulé dans limites UU (simulation niveau national puis restriction)

Enjeux

- Champ de recherche nouveau
- Pouvoir d'inférence dépend de la qualité de la combinaison des sources
- Nécessite un fort investissement méthodologique
  - Données ne sont pas produites pour une exploitation statistique
  - Assurer qualité de la combinaison avec données administratives
- Contribution:
  - Combiner données de téléphonie et données traditionnelles
  - Proposer une méthodologie pour assurer robustesse de la combinaison
  - Décrire évolution ségrégation à une échelle spatiale et temporelle fine

Données

Données de téléphonie mobile

Comptes rendus d'appels (CDR)

- Données Orange Septembre 2007, 18.5 millions de carte SIM
  - Appels et SMS: 3 milliards d'événements (France métropolitaine)
  - Géolocalisation au niveau antenne relais (présence exacte inconnue)
- Utilise pas la dimension des interactions
  - Futur travail sur ségrégation sociale

	mean	s.d.	min	P10	P25	median	P75	P90	max
Average number of daily events by user over the month	6.67	8.71	1	1.60	2.39	4.09	7.90	14.24	6260
Number distincts days phone users appear	19.98	9.16	1	5	13	23	28	30	30
Number of observations:				3,024,884,663					
Number of unique phone users:				18,541,440					

Table 1: Septembre 2007 Call Details Record: summary statistics

CDR: dimension temporelle

- Pas une trace continue
  - Individu moyen: détecté 7 fois par jour
  - Hétérogénéité forte des comportements
- Utilisation inégale du téléphone selon l'heure

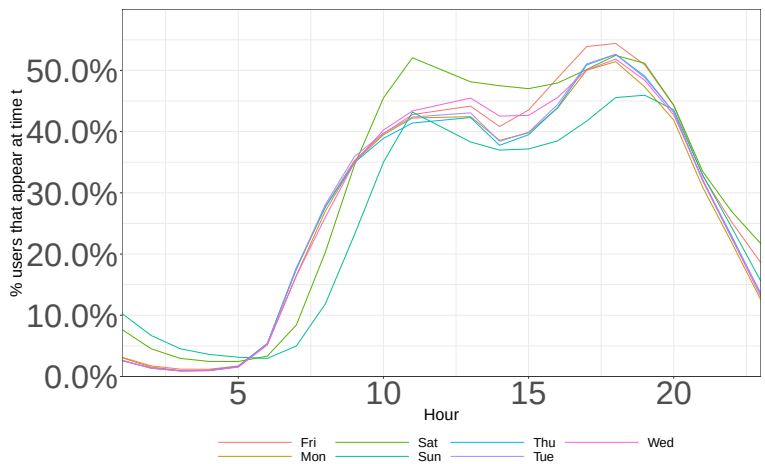


Figure 3: Utilisateurs détectés

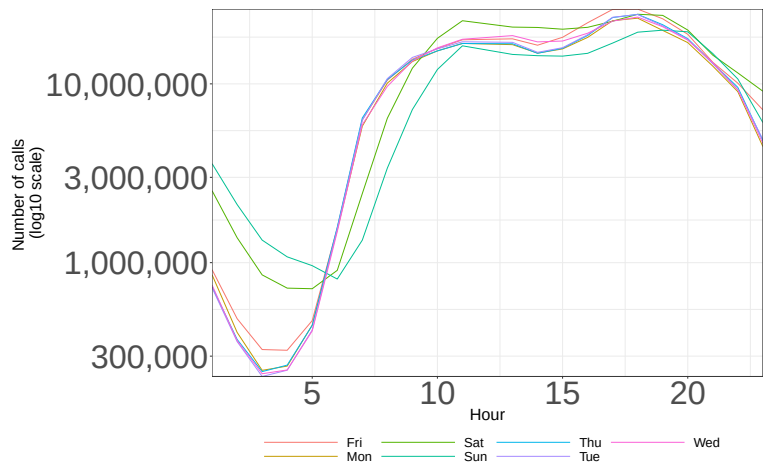


Figure 4: Appels (log)

CDR: dimension spatiale

- $\approx$ 18 000 antennes à l'échelle nationale (4000 à Paris)
- Répartition non homogène des antennes (niveau des observations)
- Privilégier analyse urbaine

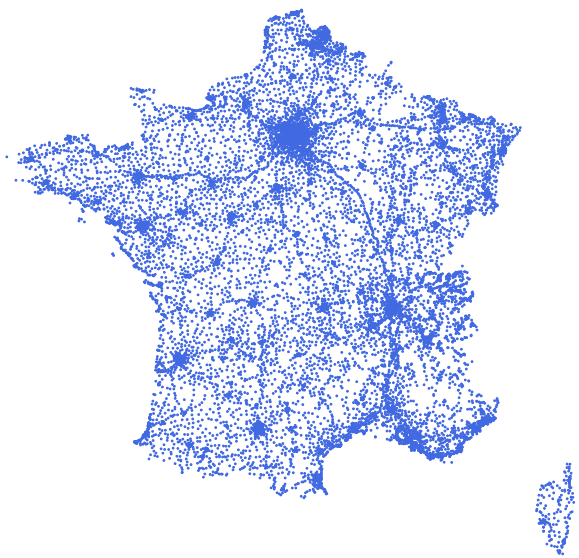


Figure 5: Répartition nationale des antennes

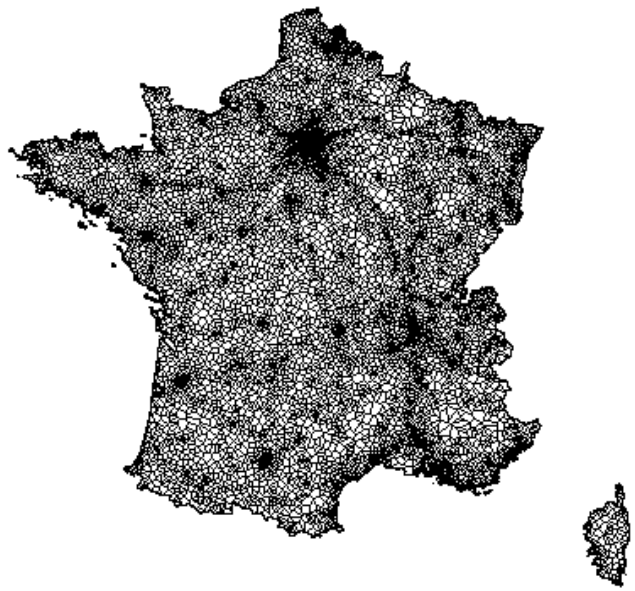
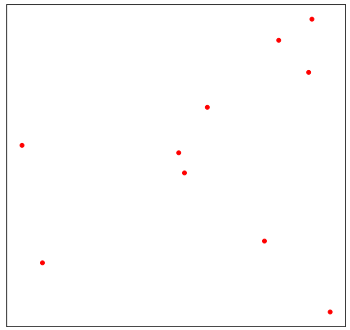
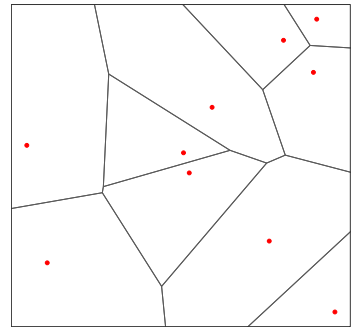


Figure 6: Aires de couverture estimées

Granularité spatiale



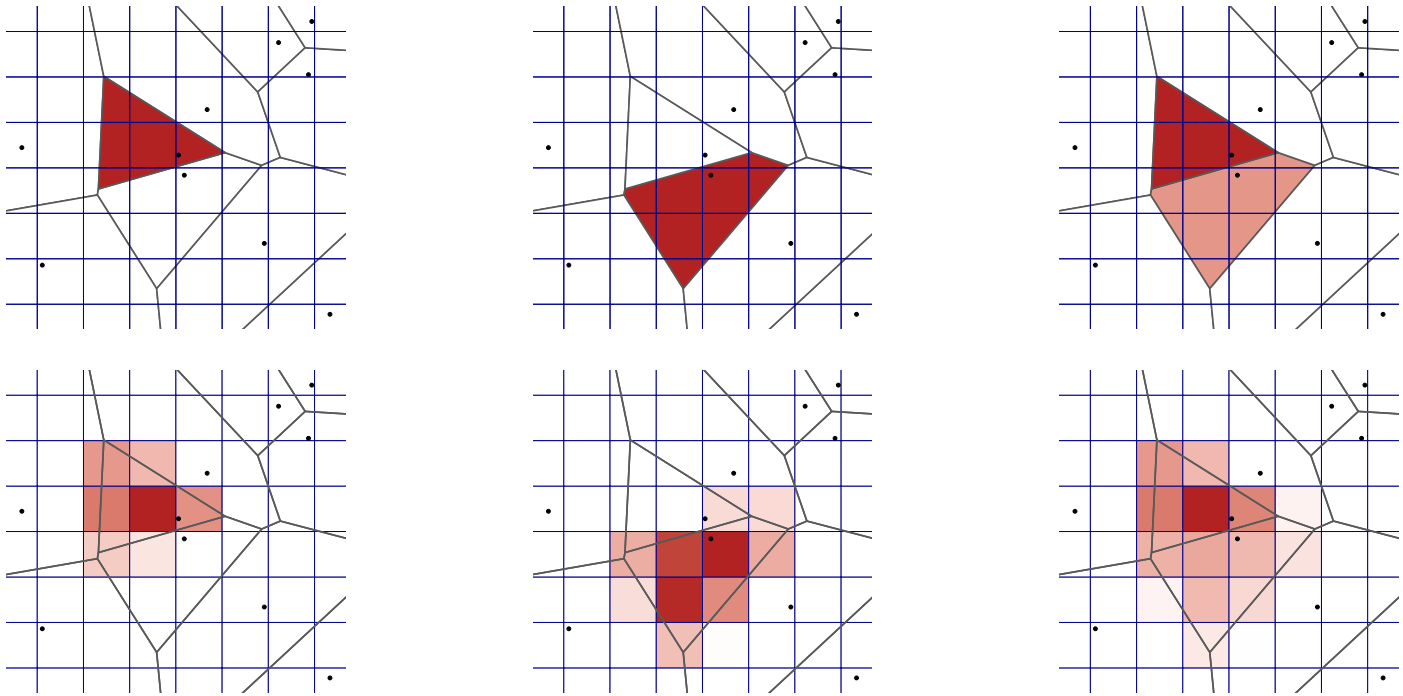


- Littérature ne distingue pas:
  - Aire de couverture: passage d'observations au niveau antenne à une aire de présence
  - Unité statistique: niveau des agrégats économiques considérés
- Modèle couverture: tesselation de Voronoi
  - Chaque point espace relié à l'antenne la plus proche
  - En l'absence d'informations sur couverture effective, simplification difficile à éviter
- Cependant, ne doit pas être l'unité statistique d'analyse
  - Dépend de la densité locale d'antennes
  - Partition espace trop hétérogène

Probabilisation de la présence

- S'abstraire du voronoi en probabilisant la présence au niveau de carreaux de taille fixe
  - Sachant que l'appel a été transmis par l'antenne v_j, quelle est la probabilité que l'individu soit présent dans un lieu donné c_i?
  - Passage par règle de Bayes Détails

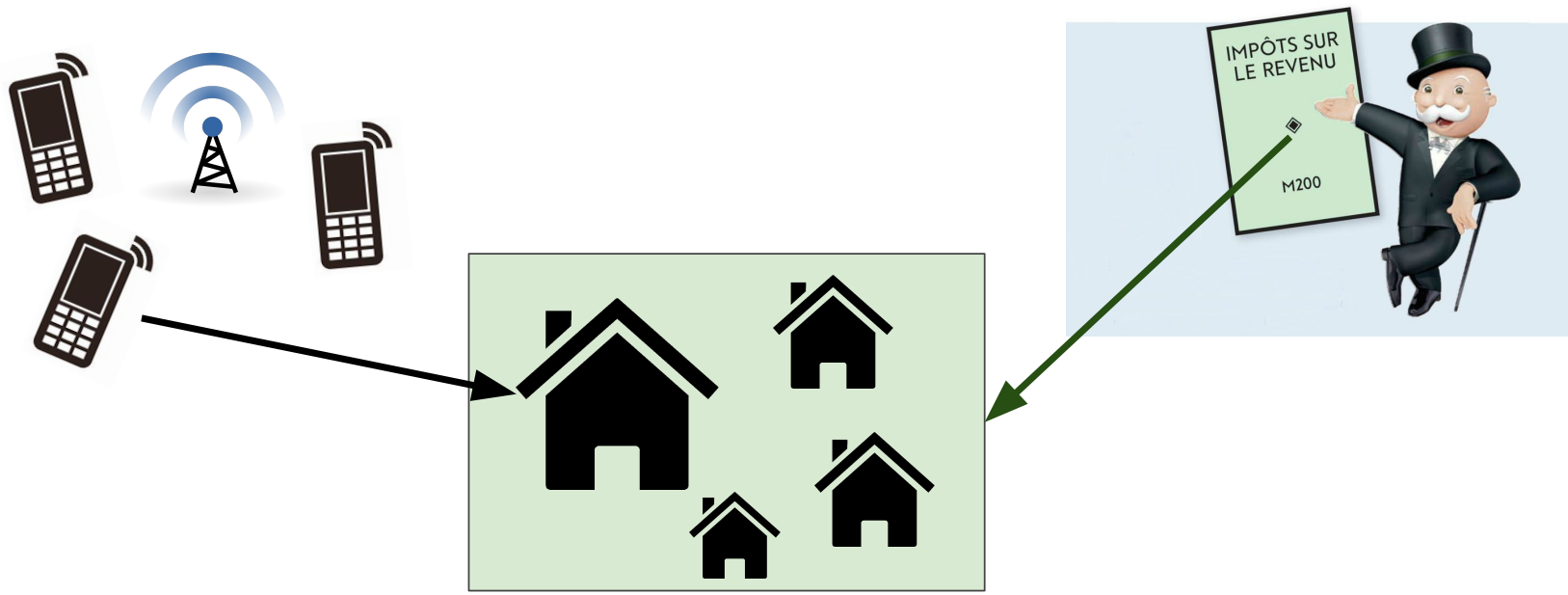
- Carreaux de 500m
  - Téléphonie: probabilisation présence & domicile au carreau
  - Filosofi: agrégats locaux sur cette grille



Données fiscales

Principe

- Caractériser utilisateurs de téléphone à partir lieu de vie
- Probabilité d'appartenir au premier/dernier décile en fonction de la distribution du revenu observée dans données fiscales



Données Filosofi

- Données fiscales géolocalisées (x, y)
  - Revenu par unité de consommation
- Ségrégation économique sur critère de revenu
  - On s'intéresse aux extrêmes de la distribution du revenu (premier et dernier déciles)
  - Définition relative du revenu: est-on plus riche/pauvre que les personnes vivant dans la même UU ?
- Approche bimodale: décompose population en classes exclusives
  - Premier décile vs reste
  - Dernier décile vs reste

Principe de la combinaison

- Fréquence d'apparition d'une sous-population (premier ou dernier décile) dans le carreau
- Agrégation spatiale au niveau du carreau i

$$p_i^{D1} = \mathbb{P}(y_x < \mu^{D1}) = \mathbb{E}(\mathbf{1}_{\{y_x < \mu^{D1}\}}) = \frac{1}{n_i} \sum_{x=1}^{n_i} \mathbf{1}_{\{y_x < \mu^{D1}\}}$$

$$p_i^{D9} = \mathbb{P}(y_x > \mu^{D9}) = \mathbb{E}(\mathbf{1}_{\{y_x > \mu^{D9}\}}) = \frac{1}{n_i} \sum_{x=1}^{n_i} \mathbf{1}_{\{y_x > \mu^{D9}\}}$$

- Si $p_i > 0.1$, sur-représentation de la sous-population dans le carreau
- Cette fréquence observée sert à simuler la sous-population d'appartenance des utilisateurs de téléphone vivant dans c_i

Méthodologie

Principe

- Le statut de l'utilisateur de téléphone est simulé à partir de son profil d'appel (seule information individuelle) et caractéristiques de Filosofi
- 3 étapes pour estimer la dynamique de la ségrégation:

1. Estimation du domicile:
   - Estimation de probabilités de résidence à partir trace d'appel en soirée: 19h-9h
2. Tirage d'un domicile et d'un revenu
   - Simulation du domicile, sachant ces probabilités de domicile
   - Simulation du revenu, à partir fréquences calculées dans Filosofi
3. Calcul d'indices de ségrégation
   - Dépendent des présences à un instant $t \Rightarrow$ vision dynamique

1. Estimation du domicile

- Domicile probabilisé à partir trace d'appel en soirée (19h-9h)
- L'ensemble de l'espace métropolitain n'est pas résidentiel
  - Dans une aire de couverture d'antenne, *a priori* sur les carreaux où trouver un espace résidentiel est le plus probable
  - A priori à partir de la densité du bâti résidentiel dans le carreau (`BD Topo`)
- Loi *a priori* est une repondération des probabilités de résidence

$$\mathbb{P}_x(c_i^{\text{home}}|v_j) \propto \underbrace{\mathbb{P}(c_i^{\text{home}})}_{\substack{\text{a priori par} \\ \text{BD Topo}}} \underbrace{\mathbb{P}_x(v_j|c_i)}_{\substack{\text{ratio surfaces:} \\ \frac{s(v \cap c)}{s(c)}}}$$

- Obtient une séquence des probabilités de domicile: $\nu_x^{\text{home}}(c_i)$ Définition
  - Utilisée pour simuler le domicile de x

2. Simulation de domicile et revenu

4 méthodes de simulation domicile pour tester robustesse estimateurs de ségrégation du groupe économique g (premier ou dernier décile)

Méthode	Domicile de x
Méthode principale	Tirage à partir ensemble probabilités de résidence ν_x^{home}
`One stage simulation`	Choix probabilité maximale de résidence: $c_i = \arg\max_{c_i} \nu_x^{\text{home}}(c_i)$
`cell_max_proba`	Domicile de x fixé par la probabilité d'être membre du groupe g maximale
`cell_min_proba`	Domicile de x par la probabilité d'être membre du groupe g minimale

Deux dernières méthodes: évaluer effet sur ségrégation de sur- ou sous-estimer la part du sous-groupe g dans la population

3. Indices de ségrégation

- Construction deux journées typiques: 24 heures de semaine, 24 heures le weekend
- Probabilités individuelles de présence dans le carreau sur la plage temporelle notée $\mathbb{P}_x(c_{it})$ Détails
- Indice de dissimilarité (Duncan & Duncan, 1955) adapté à la probabilisation de la présence:

$$ID_t^g = \frac{1}{2} \sum_{c \in \mathcal{C}} \left| \frac{\sum_{x \in \mathcal{X}} \mathbb{P}_x(c_{it}) \mathbf{1}_{x \in g}}{\underbrace{\sum_{x \in \mathcal{X}} \mathbf{1}_{x \in g}}_{\text{Number people of income group } g \text{ that are observed at time } t}} - \frac{\sum_{x \in \mathcal{X}} \mathbb{P}_x(c_{it}) \mathbf{1}_{x \notin g}}{\underbrace{\sum_{x \in \mathcal{X}} \mathbf{1}_{x \notin g}}_{\text{Number people not in income group } g \text{ that are observed at time } t}} \right|$$

- Indice classique:

$$ID = \frac{1}{2} \sum_{c \in \mathcal{C}} \left| \frac{w_c}{W_T} - \frac{n_c - w_c}{N_T - W_T} \right|$$

Résultats

Dynamique de la ségrégation

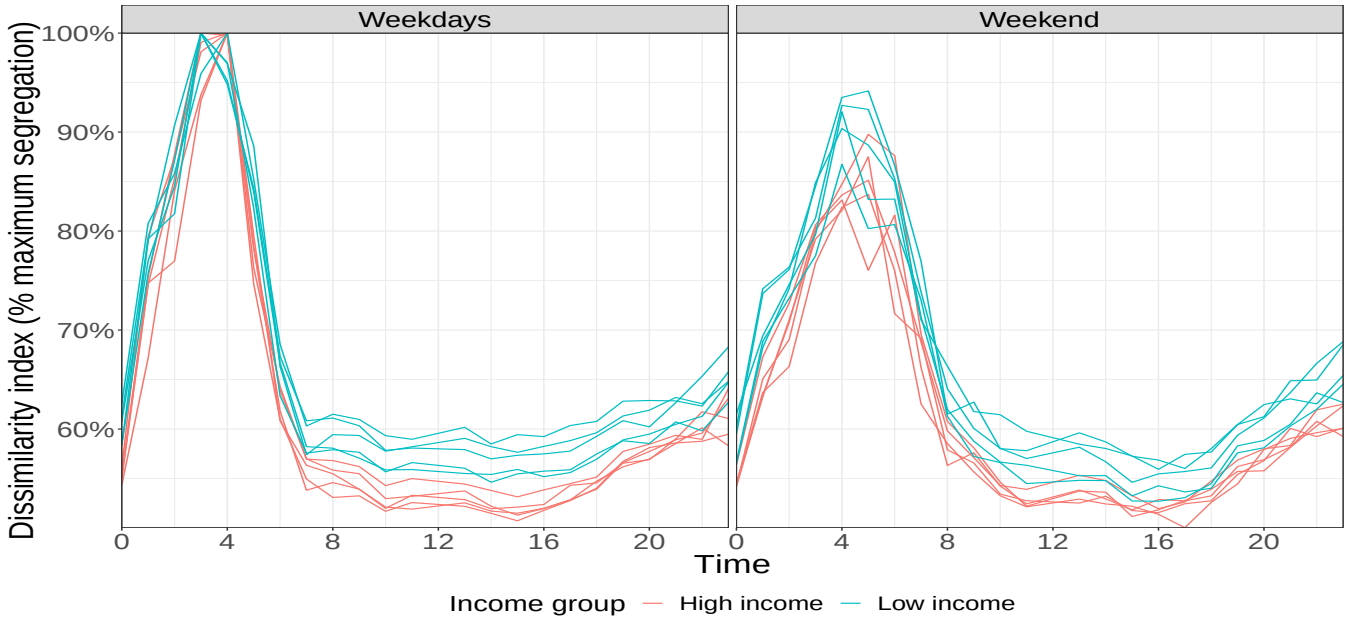


Figure 7: Marseille

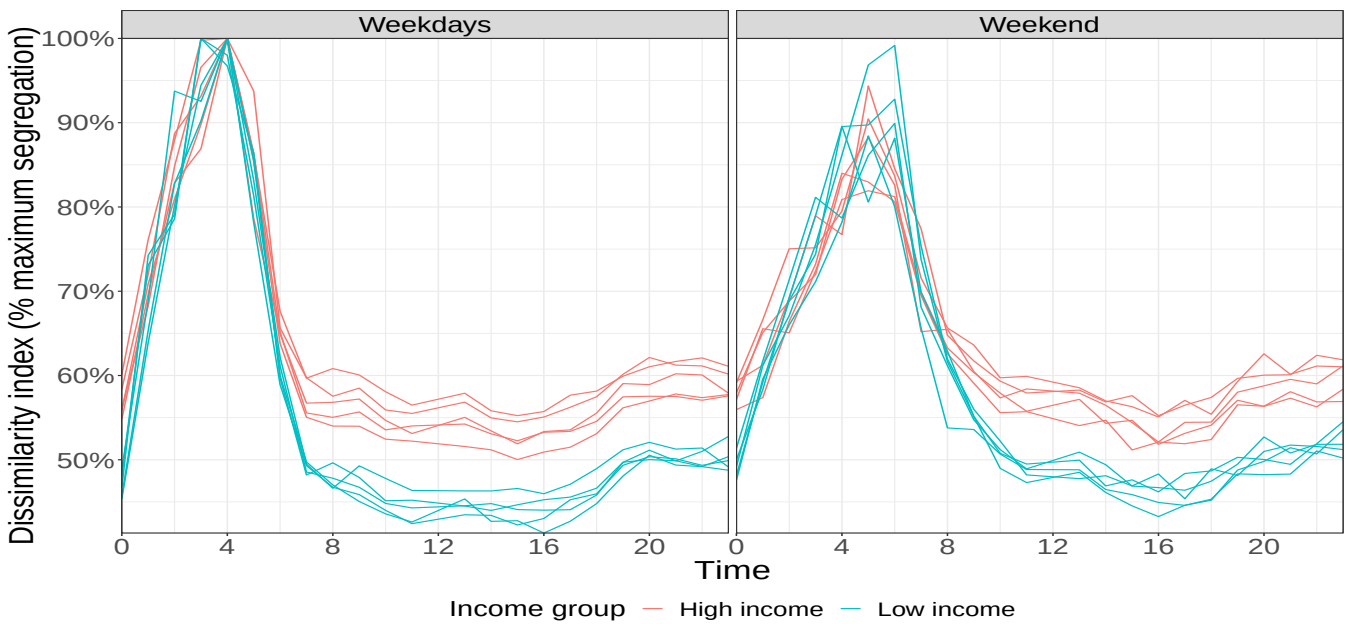


Figure 8: Lyon

Robustesse: résultats pour Marseille

Resultats pour Lyon

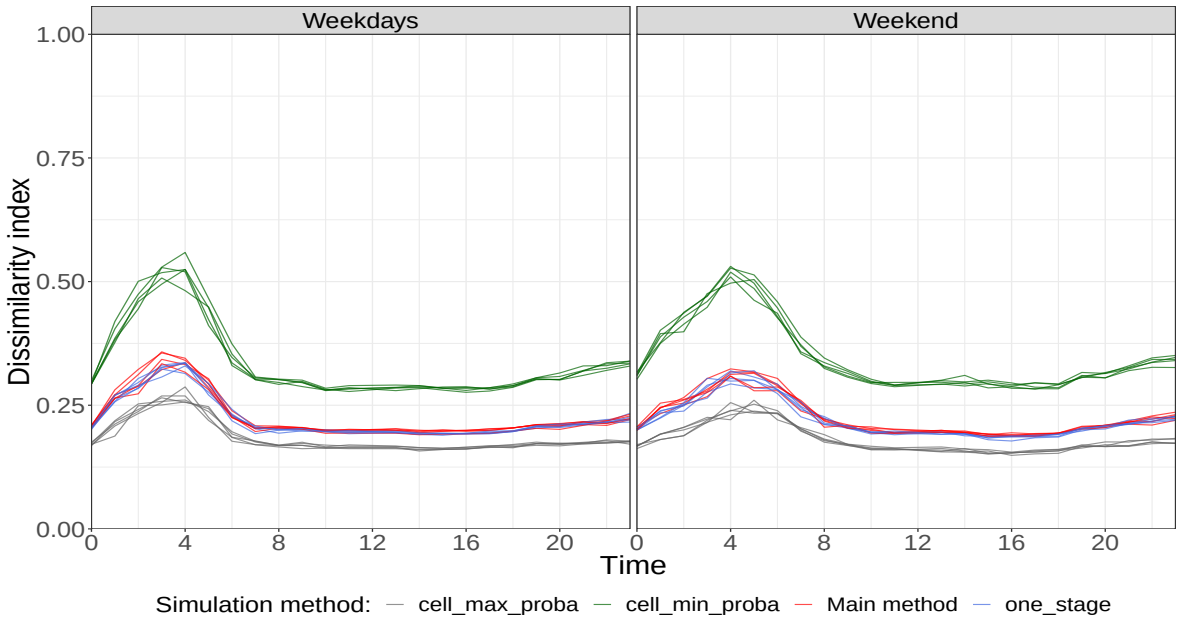


Figure 9: Premier décile

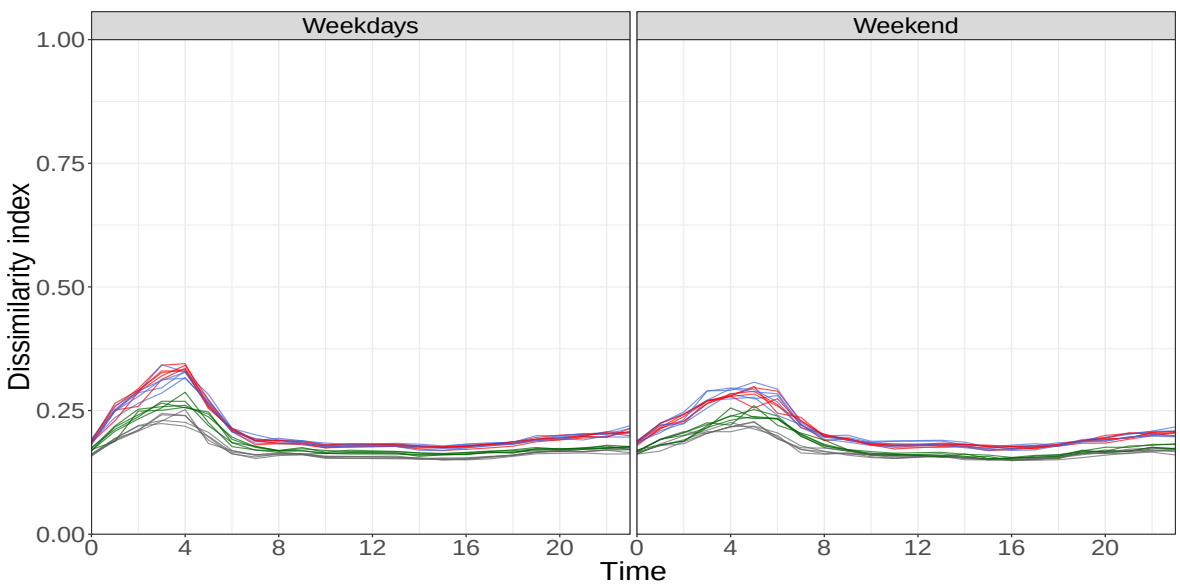


Figure 10: Dernier décile

Comparaison aux niveaux Filosofi

- Ségrégation plus faible que dans `Filosofi`
  - Pas à cause de l'approche par simulation Bootstrap Filosofi
- 2 possibilités:
  1. Ne pas interpréter les niveaux mais la dynamique
  2. Utilisateurs manquants pendant la nuit: supposer qu'ils sont chez eux

	Marseille		Lyon	
	Low-income	High-income	Low-income	High-income
Dissimilarity index in tax data	0.44	0.45	0.36	0.47
Max. dissimilarity index in phone data	0.34	0.34	0.26	0.28
Difference	0.1	0.11	0.10	0.19

Recalage

Imputation des utilisateurs non observés pendant la nuit

- Tous les utilisateurs ne sont pas observés à l'instant t

$$ID_t^g = \frac{1}{2}\sum_{c\in\mathcal{C}}\left|\frac{\sum_{x\in\mathcal{X}}\mathbb{P}_x(c_{it})\mathbf{1}_{x\in g}}{\underbrace{\sum_{x\in\mathcal{X}}\mathbf{1}_{x\in g}}_{\substack{\text{Number people of income group } g \\ \text{that are observed at time } t}}} - \frac{\sum_{x\in\mathcal{X}}\mathbb{P}_x(c_{it})\mathbf{1}_{x\notin g}}{\underbrace{\sum_{x\in\mathcal{X}}\mathbf{1}_{x\notin g}}_{\substack{\text{Number people not in income group } g \\ \text{that are observed at time } t}}}\right|$$

  - Quel effet sur les indices ?

- Imputer individus au domicile avec proba 1 lorsqu'ils sont manquants:
  - Imputation la nuit (19h-09h)

Imputation des utilisateurs non observés pendant la nuit

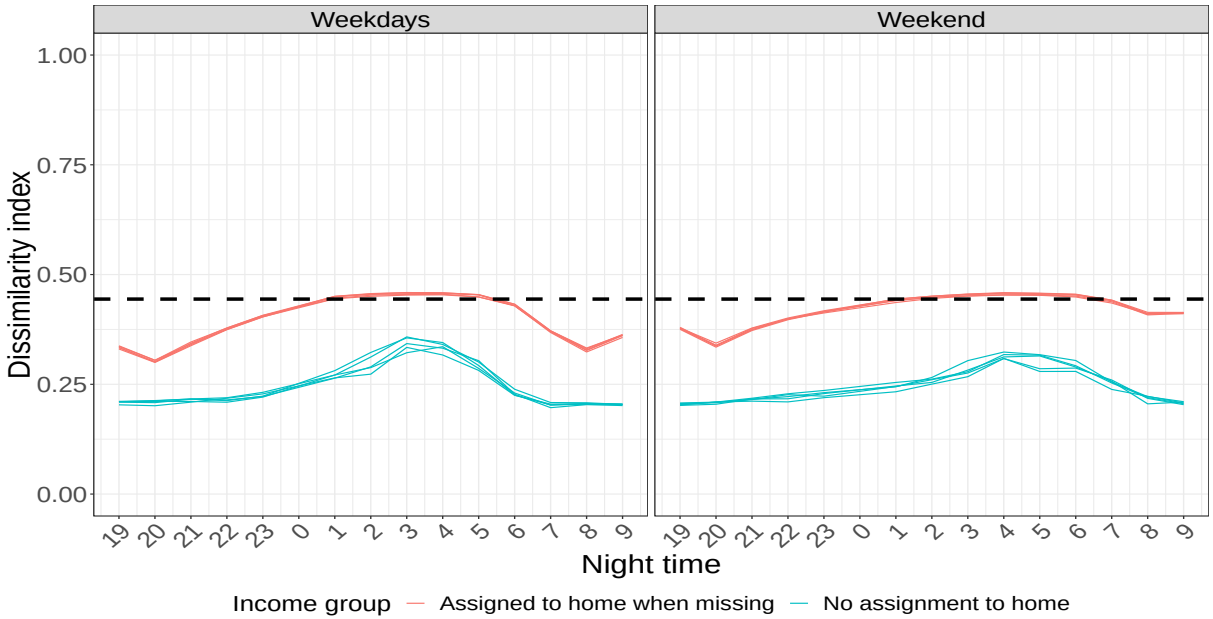


Figure 11: Premier décile

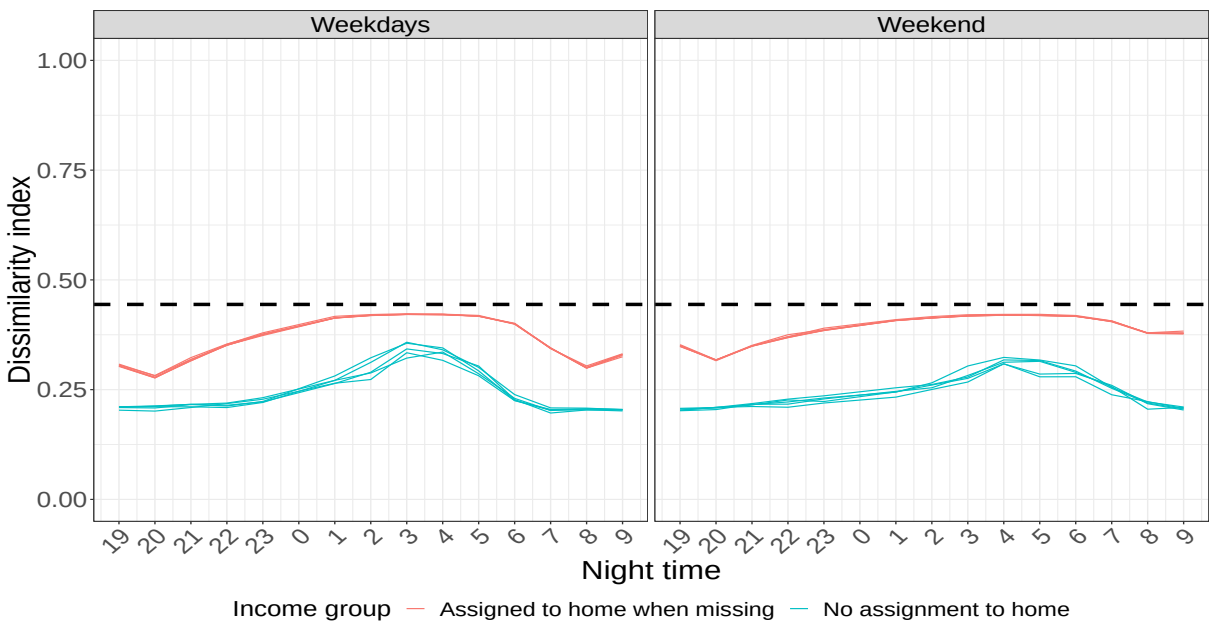
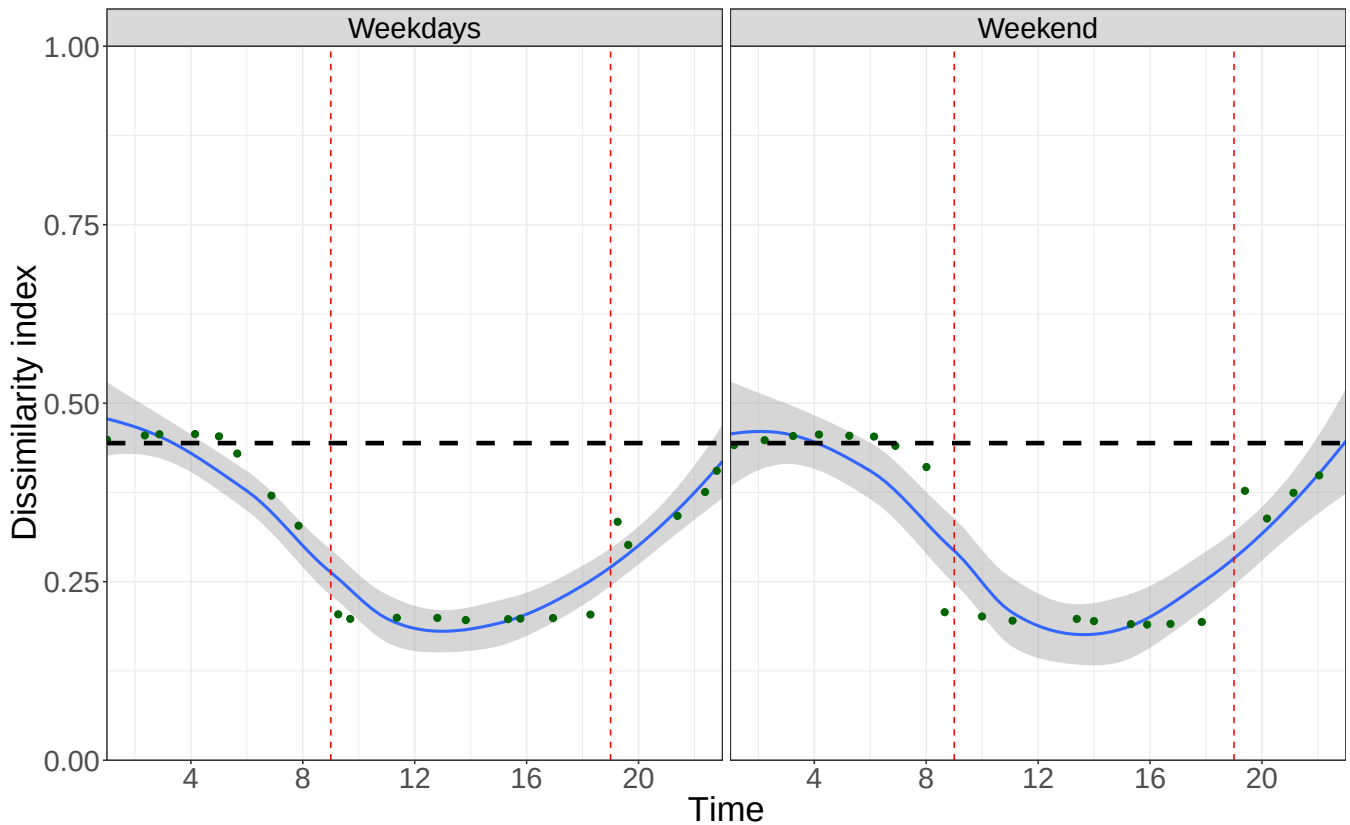


Figure 12: Dernier décile

Imputation des utilisateurs non observés pendant la nuit

- On retrouve niveau de ségrégation cohérent avec données fiscales
- Comment traiter les heures de la journée ?



Conclusion

Conclusion

- Proposition d'une méthode pour combiner données de téléphonie et données administratives
- Premiers résultats à confirmer:
  - Ségrégation moins marquée pendant la journée que le suggère la ségrégation résidentielle (cohérent avec Le Roux et al., 2017)
  - Pas de différence marquée dans la dynamique entre premier et dernier déciles
  - Recalage permet d'avoir un niveau de ségrégation cohérent avec `Filosofi`
- Suite du travail:
  - Investissement méthodologique à approfondir
  - Généraliser à d'autres villes
  - Comprendre la dynamique de la ségrégation à une échelle infra-urbaine Exemple

Appendix

Annexe méthodologique

Présence au carreau

- La probabilité qu'un événement mesuré dans l'antenne v_j à l'instant t ait lieu dans le carreau c_i est égal à

$$p_i^j := \mathbb{P}(c_i|v_j) = \frac{\mathbb{P}(c_i \cap v_j)}{\mathbb{P}(v_j)} = \frac{\mathcal{S}(c_i \cap v_j)}{\mathcal{S}(v_j)}$$

- La probabilité d'être présent à l'instant t dans le carreau c_i (on note cette double condition c_{it}) est la recollection des probabilités conditionnelles

$$\forall c_{it} \in \mathcal{C}, \quad \mathbb{P}_x(c_{it}) = \sum_{v_{jt} \in \mathcal{V}} \mathbb{P}(c_{it}|v_{jt}) \mathbb{P}_x(v_{jt}) \tag{1}$$

avec $\mathcal{V}$ ensemble des antennes/voronois et $\mathcal{C}$ cellules de 500m.

Estimation du domicile

- Détection domicile uniquement: événement au niveau du voronoi repondéré selon la formule suivante

$$\mathbb{P}_x(c_i^{\text{home}}|v_j) \propto \underbrace{\mathbb{P}(c_i^{\text{home}})}_{\substack{\text{a priori par} \\ \text{BD Topo}}} \underbrace{\mathbb{P}_x(v_j|c_i)}_{\substack{\text{ratio surfaces:} \\ \frac{s(v \cap c)}{s(c)}}}$$

- Domicile estimé de l'individu x au niveau du carreau c_i en sommant tous les événements mesurés au niveau des voronoi v_j:

$$\nu_x^{\text{home}}(c_i) = \frac{1}{\alpha_x} \sum_{v \in \mathcal{V}} \mathbb{P}_x(c_i^{\text{home}}|v_j)\mathbb{P}(v_j)$$

- avec α_x un terme de normalisation pour avoir $\sum_{c_i} \nu_x^{\text{home}}(c_i) = 1$

Annexe résultats

Robustesse: Lyon

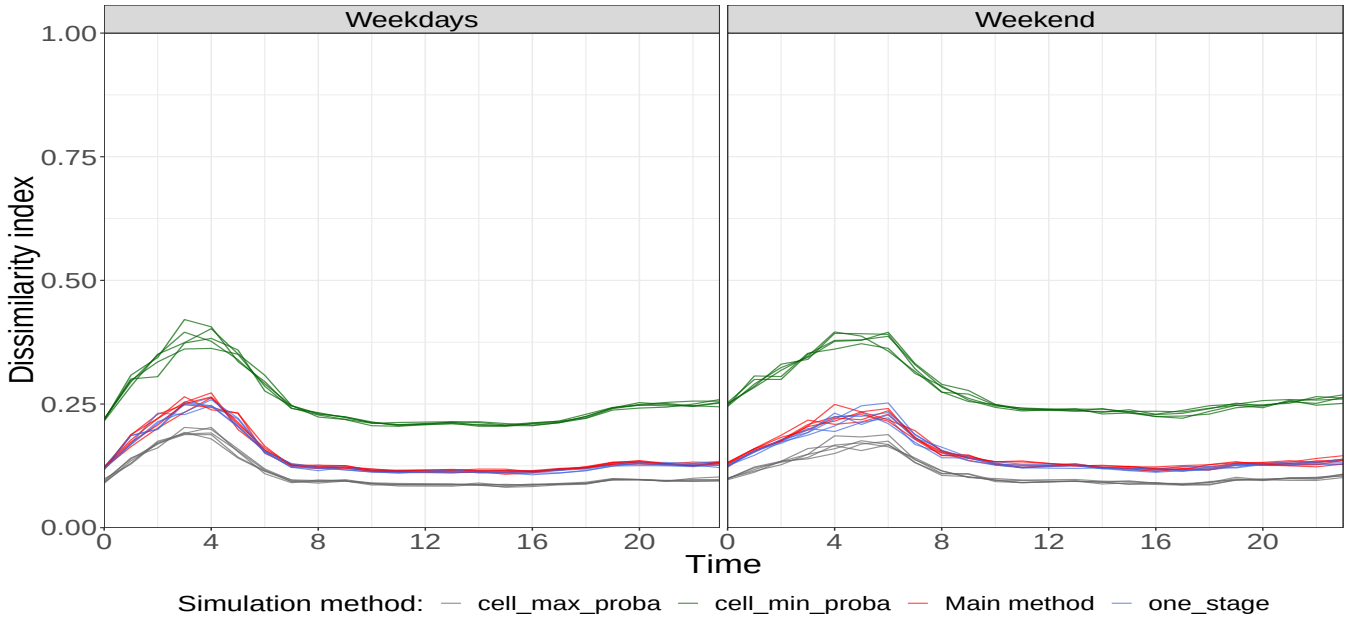


Figure 13: Premier décile

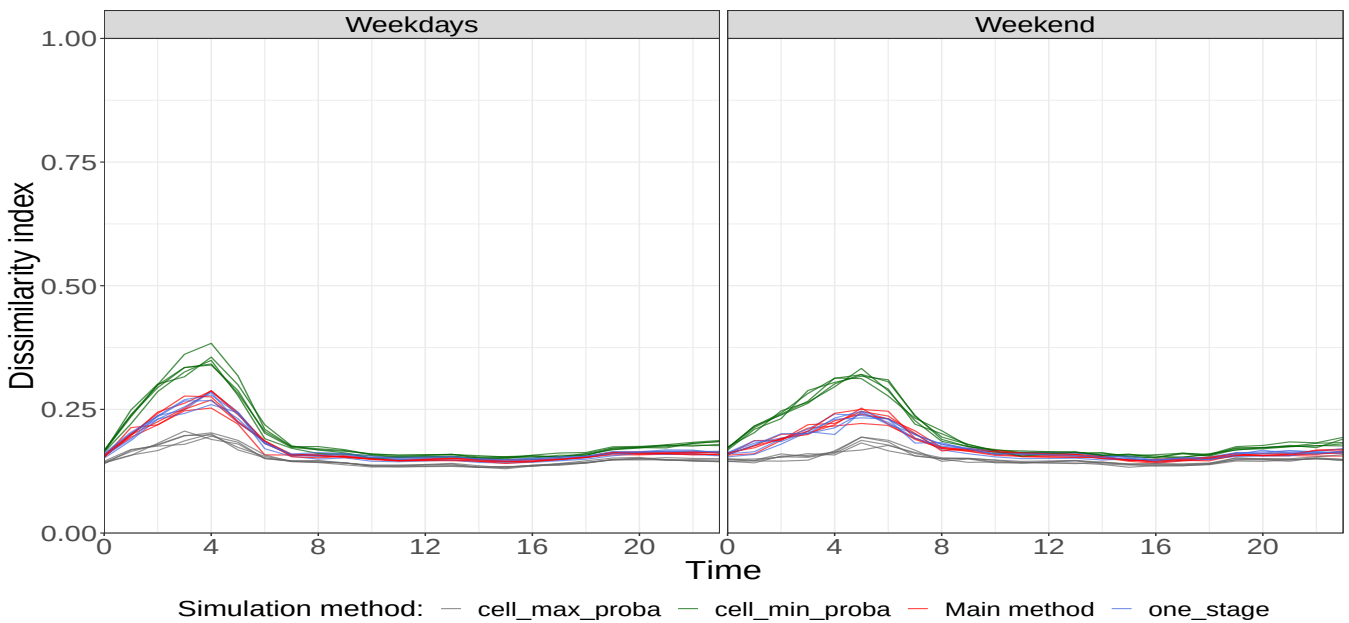


Figure 14: Dernier décile

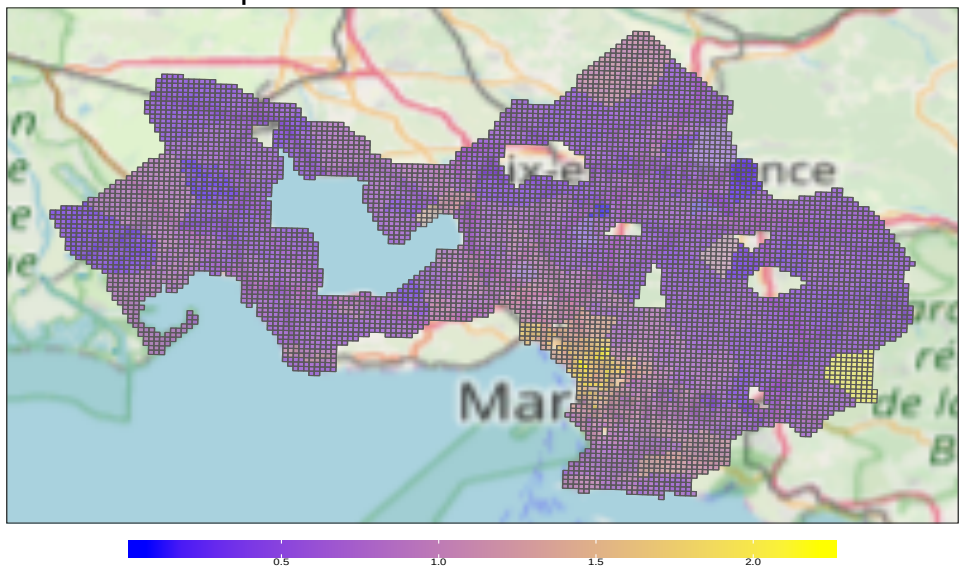
Effet simulation sur ségrégation: bootstrap données fiscales

	Dissimilarity Index			
	Observed value	*Bootstrap design*		
		(1)	(2)	(3)
		Marseille		
Low-income	0.4441	0.4439	0.446	0.457
		[0.4422;0.4455]	[0.4429;0.4502]	[0.4515;0.4611]
High-income	0.4536	0.4088	0.415	0.4226
		[0.4070;0.4107]	[0.4117;0.4186]	[0.4187;0.4269]
		Lyon		
Low-income	0.3584	0.359	0.3626	0.352
		[0.3572;0.3606]	[0.3602;0.3654]	[0.3483;0.3561]
High-income	0.4691	0.3969	0.4021	0.4028
		[0.3944;0.3987]	[0.3986;0.4059]	[0.3985;0.4067]

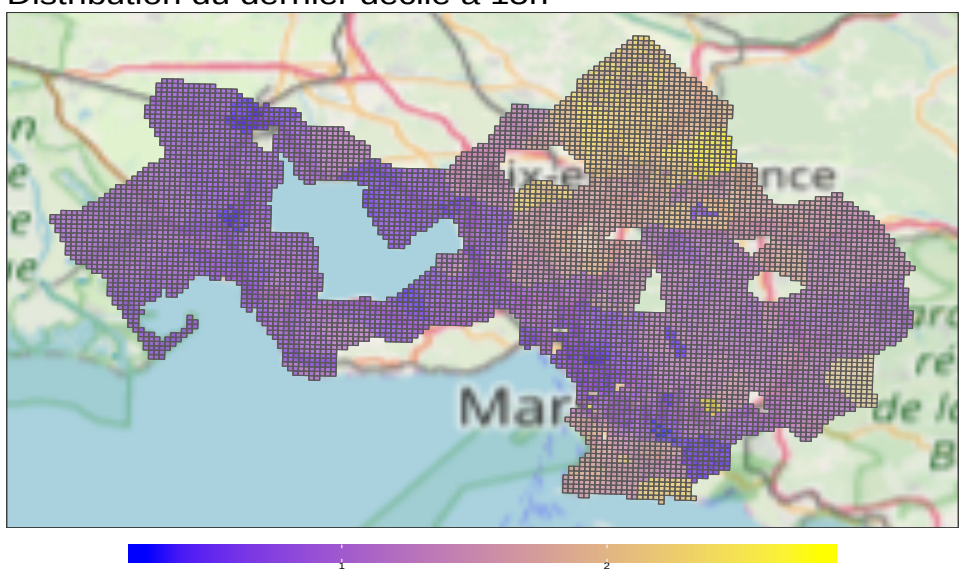
Notes:
Median dissimilarity index over 100 iterations is reported. 95% confidence intervals are reported into brackets
(1): Bootstrap inside each cell with uniform weights (probability being chosen: $1/n_i$)
(2): Bootstrap inside each cell with uniform weights for 1/3 population (probability being chosen: $\frac{1}{3n_i}$)
(3): Bootstrap inside each cell with uniform weights for population from mobile phone data (probability being chosen: $\frac{1}{n_{\text{phone}}}$).

Cartographie de la ségrégation à 15 heures

Distribution du premier décile à 15h

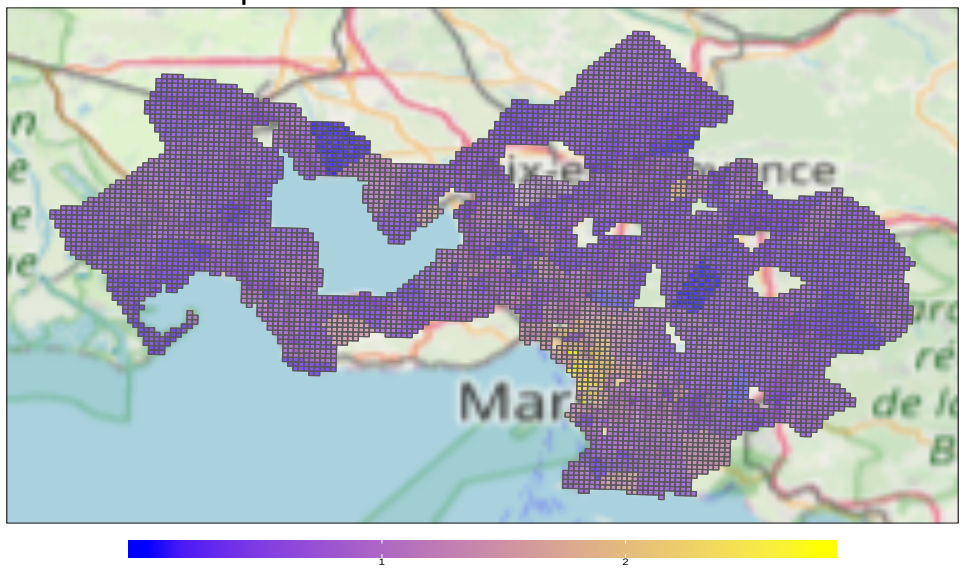


Distribution du dernier décile à 15h



Cartographie de la ségrégation à 22 heures

Distribution du premier décile à 22h



Distribution du dernier décile à 22h