Estimation de la population à partir des metadata du trafic du réseau mobile

Ghazaleh Khodabandelou[1], Vincent Gauthier[1], Mounim El-Yacoubi[1], Marco Fiore[2]

[1]SAMOVAR, Telecom SudParis, CNRS, University Paris-Saclay

[2]CNR-IEIIT, Italy

RESUME

Les Smartphones et autres appareils mobiles sont aujourd'hui omniprésents dans la société. Une des conséquences de l'ubiquité des communications mobiles est que les opérateurs de réseaux mobiles peuvent désormais facilement recueillir une quantité immense de données à haute résolution sur les comportements des grandes populations d'utilisateurs. Les informations extraites des données de trafic du réseau mobile sont très pertinentes dans le contexte de la cartographie de la population :elles fournissent un outil efficace pour l'estimation automatique et directe des densités de population, ce qui permet de surmonter les limites des sources de données traditionnelles telles que les recensements et les enquêtes. Dans cet article, nous proposons une nouvelle approche pour estimer la densité de population à l'échelle urbaine, à partir des métadonnées agrégées du trafic du réseau mobile. Notre approche permet d'estimer les populations à la fois statiques et dynamiques, et permet d'obtenir une amélioration significative en termes de précision par rapport aux solutions de l'état de l'art.

Mots clés : Smartphones, Réseaux mobiles, Estimation de la densité de population.

Introduction

Les données du trafic des réseaux mobilesoffrent le potentiel d'automatiser l'estimation en temps quasi réel de la densité de population,grâce à la relation classiqueentre le volume d'activités mobiles et cette dernière [9].Nous introduisons une nouvelle approche pour l'estimation de la population, évaluéeavec plusieurs jeux des données de signalisation mobiles. Les résultats obtenusmontrent que notre modèle atteint une corrélation nettement améliorée avec les données de réalité-terrain, typiquement dans la gamme 0,80 à 0,87. De plus :

- Nous concevons une solution exclusivement basée sur les métadonnées collectées par l'opérateur des réseaux mobiles, évitant ainsi le recours à des données complexes et mélangées ;
- Nous montrons que les données de présence des abonnés inférées à partir de leurs communications mobilessontun meilleur proxy de la distribution de la population par rapport à des mesures adoptées précédemment;
- Nous introduisons un certain nombre de filtres sur les données originales qui permettent d'affiner les estimations de la répartition de la population ;
- Nous évaluons notre méthodologie dans de multiples scénarios urbains, en obtenant toujours de bons résultats ;
- Nous dévoilons la relation multivariée entre la densité de la population, la présence des abonnés et leur niveau d'activités ;
- Nousutilisons notre modèle pour générer des représentations dynamiques de répartition de la population.

1 Données

Dans ce travail, nous misons sur plusieurs jeux de données mis à disposition par Telecom Italia Mobile (TIM) dans leur 2015 Big Data Challenge [15]. Plus précisément, nous nous concentrons sur trois grandes zones urbaines pour lesquelles des données substantiellessont disponibles, à savoir, les agglomérations de Milan, Turin et Rome. Pour chaque ville, nous recueillons des données décrivant des activités de trafic mobiles et la distribution de la population. Nous déduisons également des informations surl'aménagement dusol (type d'activité, zone résidentielle, zone commerciale, etc.).

1.1 Trafic de réseaux mobiles

Les données de télécommunication couvrent les mois de Mars et Avril 2015, et décrivent le volume de trafic divisé par type (appels vocaux entrant/sortant, texte SMS entrant/sortant, et Internet), et par la présence des abonnés. Tous les paramètres sont regroupés dans le temps et dans l'espace. Dans le temps, les données sont totaliséessur des intervalles de 15 minutes. Dans l'espace, les métriques sont calculées sur une tessellation de grille irrégulière, dont les cellules géographiques ont des tailles allant de 255×325 m² à 2×2.5 km². Le nombre de cellules est de 1419 pourMilan, 571 pour Turin et 927 pour Rome. Alors que la voix, le texte, et les volumes de données sont directement calculés à partir de la demande enregistrée, les informations de présence sont le résultat d'un prétraitement simple, effectué par l'opérateur de réseau mobile. Fondamentalement, chaque abonné est associé à la cellule géographique où il/elle a effectué son/sa dernière action (par exemple, lancé un appel, reçu un SMS, etc.). Si la présence d'un utilisateur est enregistrée sur la place A, puis il effectue une action sur la place B au temps t1, et ensuite il interagit avec le réseau dans le carré C au temps t2,, sa présence sera

enregistrée comme suit: à t<t1, la présence du l'utilisateur est enregistrée dans A; pour t=t1, la présence de l'utilisateur est déplacée de A à B; pour l'instant t1<t<t2, la présence de l'utilisateur est enregistrée dans B; à t=t2, l'utilisateur est déplacée de B à C; pour t>t2, la présence de l'utilisateur est enregistrée en C jusqu'à ce qu'il effectue une action dans un carré différent. Fig.1 montre un exemple de prétraitement des données de présence : l'emplacement d'un utilisateur spécifique est détecté toutes les 15 minutes, s'il effectue au moins une action au cours de cet intervalle de temps.

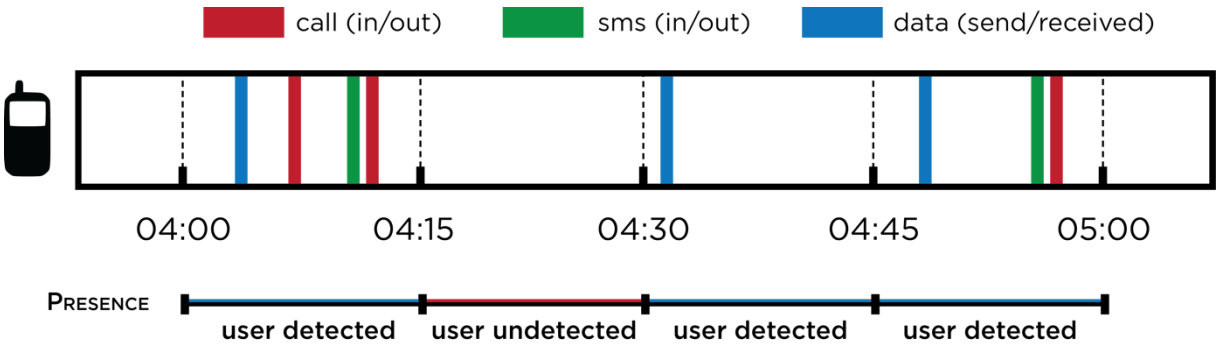

Figure 1La présence d'un abonné telle qu'elle est définie dans le TIM Big Data Challenge

1.2 Distribution de la population

Les données sur la distribution de la population proviennent du recensement fait en 2011 en Italie,effectué par l'organisation nationale pour les statistiques, ISTAT. Il comprend :lapopulation, le nombre et les caractéristiques structurelles des maisons et des bâtiments [16]. Plus précisément, la population est mesurée par familles, les personnes temporairement présentesdans le domicile.Dans notre étude, nous allons nous servir de ces données comme une vérité terrainpour la distribution de la population statique dans les régions urbaines de référence. À cette fin, nous avons besoin d'assurer la cohérence spatiale entre le trafic de réseaux mobiles et les données du recensement. Nous procédons comme suit. Admettons$U_i$comme le nombre total d'habitants dans la zone administrative*i*, et$A_j$comme la surface de la zone de couverture cellulaire*j*. La densité de la population$\rho_i$dans une cellule géographique*i*est définiesuit :

$$\rho_i = \frac{1}{A_i} \sum_{j=1}^{K} U_i \frac{A_{i \cap j}}{A_j} \qquad (1)$$

où$A_i$est la surface de la cellule*i*, *K*désigne le nombre total de zones administratives, et$A_{i \cap j}$représente la surface d'intersection de la cellule*i*et de la zone administrative*j*.

2 Estimation statique de la population

Notre modèle d'estimation de la population de référence estfondé sur des résultats précédents qui démontrent une relation de puissance constante entre l'activité du trafic des réseaux mobiles $\sigma_i$et la densité de la populationρ_id'une même région *i* [8], [9]. Ainsi :

$$\rho_i = \alpha \sigma_i^{\beta} \qquad (2)$$

Les paramètres α et β représentent le point d'intersection (ou le rapport d'échelle) et la pente (ou l'effet de ρ_i sur σ_i) du modèle. En transformant la formule à une échelle logarithmique, on obtient $\log(\rho_i) = \log(\alpha) + \beta.\log(\sigma_i)$.On peut alors utiliser un modèle de régression pour estimer les paramètres α et β dans l'équation (2).

Malheureusement, les résultats de régression - peu importe le type de données - seront médiocres si celle-ci est exécutée sur les données de télécommunicationsbrutes. Considérons les données de la ville deMilan : une visualisation simple des ces données(Fig.2) dévoile que l'ensemble des données souffre d'hétérogénéité et hétéroscédasticité. La figure illustre la densité d'appels, de SMS et de présence en fonction de la densité de population associée. La variance élevée entre les

variables dépendantes et indépendantes révèle l'absence d'homoscédasticité dans les données. Cependant, les modèles de régression classiques supposent qu'il n'y a pas d'hétéroscédasticité dans les données, et, sur la base de cette hypothèse, déduisent les meilleurs estimateurs linéaires sans biais.Il en résulte que le filtrage des données bruitées (cf. Fig. 2) est une étape nécessaire pour appliquer des modèles de régression. Dans ce qui suit, nous introduisons des techniques de débruitage des données sur plusieurs niveaux. Nous nous concentrons sur l'étude du cas de Milan, pour généraliser ensuite les résultats de notre approche en la testant dans d'autres villes.

2.1 Filtrage sur les horaires

Une deuxième dimension sur laquelle le filtrage des données est appliqué est la dimension temporelle. Commecela est montré dans la littérature [9], la corrélation entre les données de trafic sur réseaux mobiles et la densité de population varie au fil du temps (Figure 3). Le coefficient de corrélation est le plus élevé dans la nuit, dans la plage 4-5h. Ce résultat est très raisonnable, puisque les données de la population ISTAT se réfèrent au logement. Ainsi, on choisit la présence des abonnés durant cette heure précise afin d'étalonner nôtre le modèle de régression. Dès lors, on définit σ_i en (1) comme étant la présence de l'utilisateur dans la cellule *i* dans la plage 04-05 h.

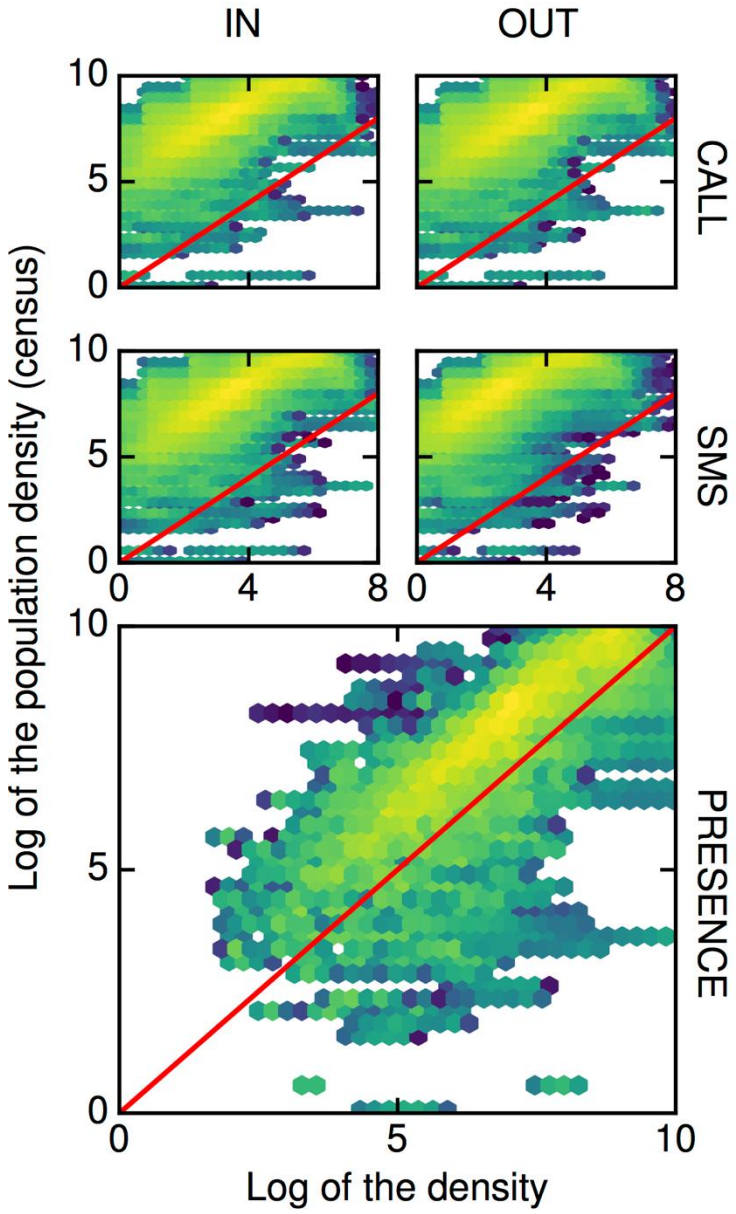

Figure 2: Milan. Densité de population ISTAT en fonction du volume d'appels, du volume de SMS, et deprésence des abonnés.

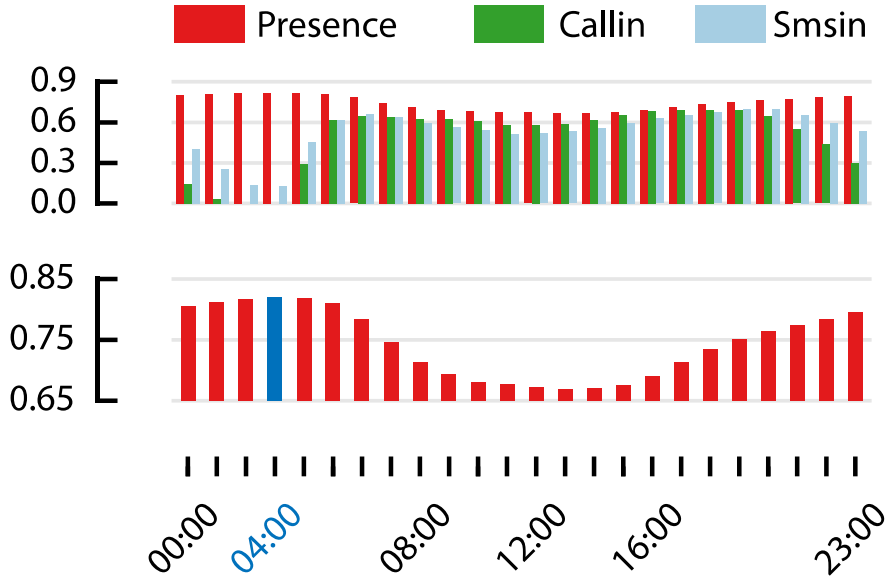

Figure 3: Milan. Haut : corrélation entre les différents types de données de trafic des réseaux mobiles et la densité de recensement de la population, sur une base horaire. Bas : zoom sur les métadonnées de présence.

2.2 Régression avec RANSAC

Afin d'estimer les paramètres α et β dans (2), nous employons le régresseur RANSAC [19] sur les données de présence filtrées. RANSAC estime les paramètres d'un modèle à partir d'observations de manière itérative, et détecte automatiquement et exclut les valeurs aberrantes (cf. Fig. 5).Les valeurs aberrantes et acceptables détectées par RANSAC dans les données sont indiquées en gris et en violet, respectivement. Les paramètres estimés sont $\hat{\alpha}$=1.265, $\hat{\beta}$= 0,979. Le résultat souligne la relation quasi-linéaire entre la présence des abonnés et la population, comme en témoigne également leurraccord linéaire (ligne noire en pointillés).

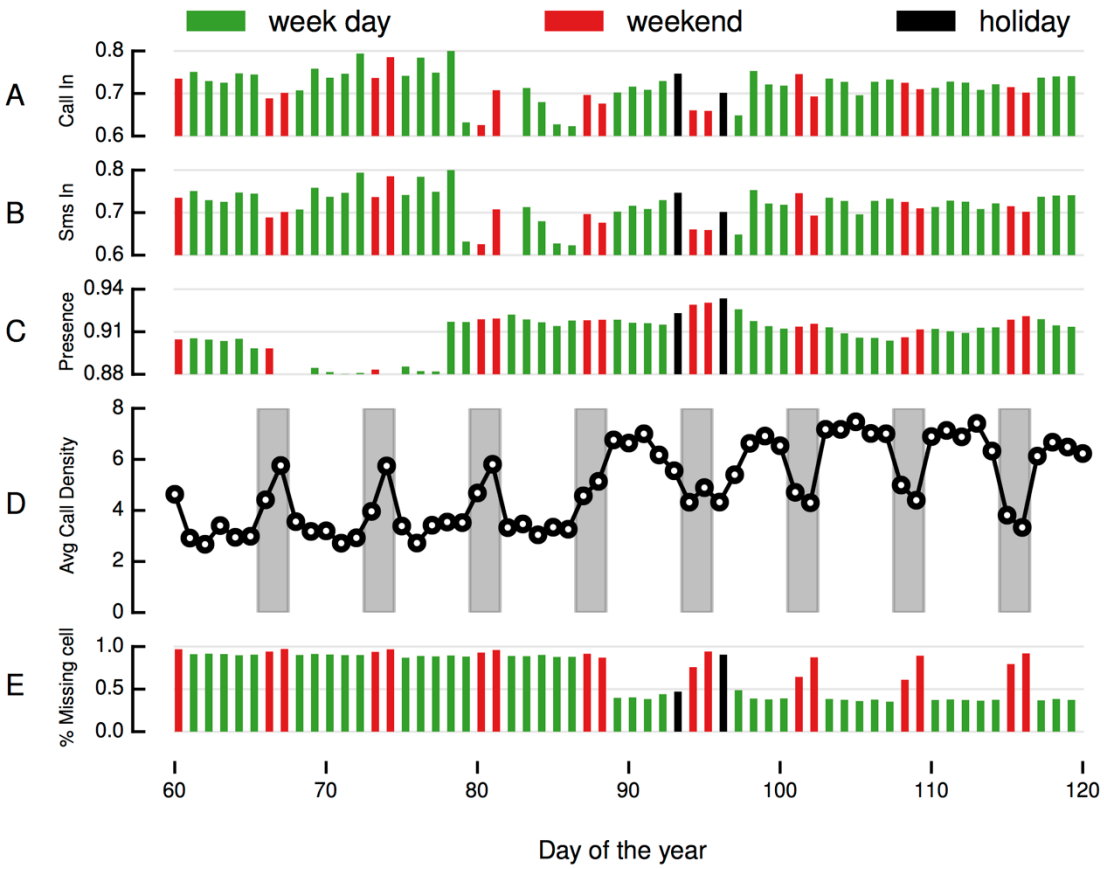

Figure 4Milan. Coefficient de corrélation de Pearson pour de différentes données de trafic des réseaux mobiles de 60 jours, en Mars et Avril 2015, 4-5 h. A) Appel entrant (allant jusqu'à 0,8), B) SMS entrant (allant jusqu'à 0,8), C) la présence

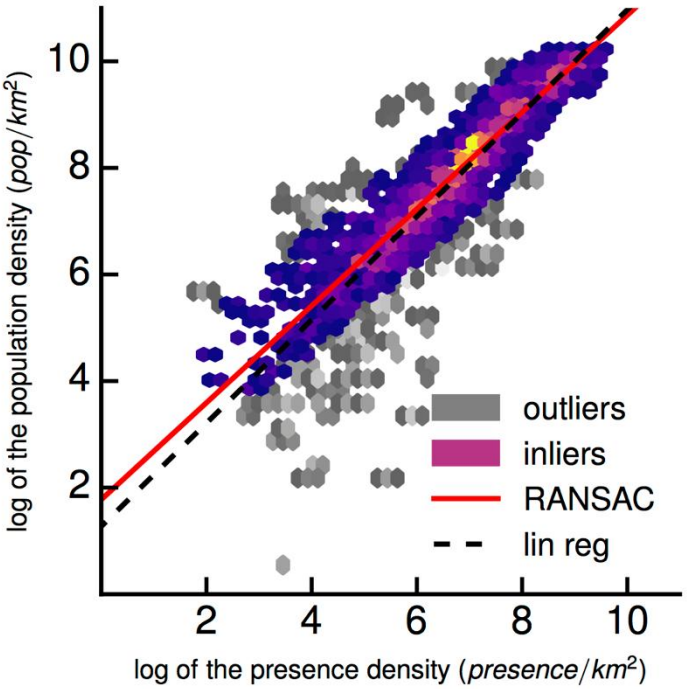

Figure 5: Milan. RANSAC et régression linéaire sur les données de présence des abonnés, filtrées.

2.3 Evaluation du modèle

Le modèle de régression permet de calculer une estimation de la densité statique de la population $\hat{\rho}_i$ en fonction des métadonnées de présenceσ_i, au sein de chaque cellule spatiale *i*,par:

$$\hat{\rho}_i = \hat{\alpha}\sigma_i^{\hat{\beta}}. \tag{3}$$

La population estimée$\hat{\rho}_i$peut alors être comparée à la densité vérité-terrain$\rho_i$issue desdonnées de recensement ISTAT. À cette fin, nous utilisons le coefficient de détermination R^2, et la Normalized Root Mean Square Error (*NRMSE*) i.e. l'erreur quadratique moyennenormalisée. R^2 fournit une mesure de la qualité de l'ajustement des estimations, alors que *NRMSE* décrit la fraction de l'erreur entre les valeurs prédites par le modèle et les valeurs des données de recensement. Le coefficient R^2 est calculé comme :

où *N* désigne le nombre de cellules dans l'espace de tessellation, et $\overline{\rho}$ est la densité moyenne calculée sur l'ensemble des cellules. La NRMSE facilite la comparaison des résultats du modèle dans différents contextes. Elle est définie par :

$$NRMSE = \frac{1}{\rho_{max} - \rho_{min}} \sqrt{\frac{\sum_{i=1}^{N}(\hat{\rho}_i - \rho_i)^2}{N}}, \tag{5}$$

où ρ_{max} et ρ_{min} sont les densités de population maximale et minimale enregistrées dans la région cible. Puisque le modèle est appris sur les données de recensement ISTAT, nous adoptons une double procédure de validation croisée, comme suit. Pour chaque test, nous séparons les données en deux sous-ensembles : les deux tiers des données sont utilisées comme un ensemble d'apprentissage et le tiers restant comme un ensemble de test. Ensuite, l'ensemble d'apprentissage est utilisé pour apprendre les paramètres du modèle, et le modèle résultant est évalué sur l'ensemble de test.

Le résultat de base, dans l'étude du cas de Milan, est représenté dans Fig. 7. Le plot du haut (A) montre le coefficient de détermination R^2 obtenu avec les données d'apprentissage et de test. Une première observation est que les résultats sont comparables pour les données d'apprentissage et de test, ce qui valide notre modèle.Fig. 7 montre les résultats de l'étude

sur Milan. On peut constater que les performances sont bonnes pour les zonesde déplacements, touristiques et commerçantes, acceptables pour les zones d'affaires et de la vie nocturne, et mauvaises pour les zones universitaires. Nous pensons que ce phénomène pourrait être dû à la présence d'activités de communications mobiles pendant la nuit dans les campus universitaires (par exemple, les fêtes ou des activités à des fins de recherche) alors que personne n'y vit réellement. Dans tous les cas, les zones universitaires ne représentent qu'une minorité négligeable de cellules spatiales, ce qui permet de conclure quenotre modèle peut être efficacement utilisé pour estimer la densité de population dans toute région urbaine.

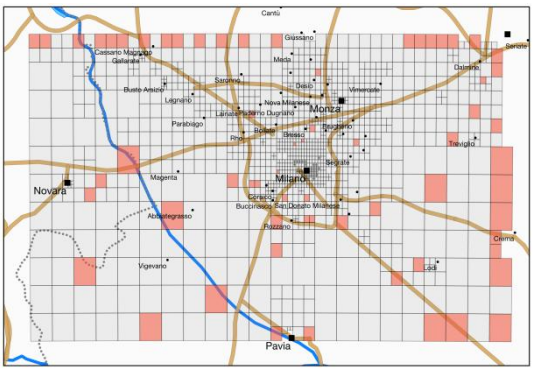

Figure 6 Milan. Répartition géographique des cellules qui le plus souvent entraînent des données aberrantes détectées par RANSAC

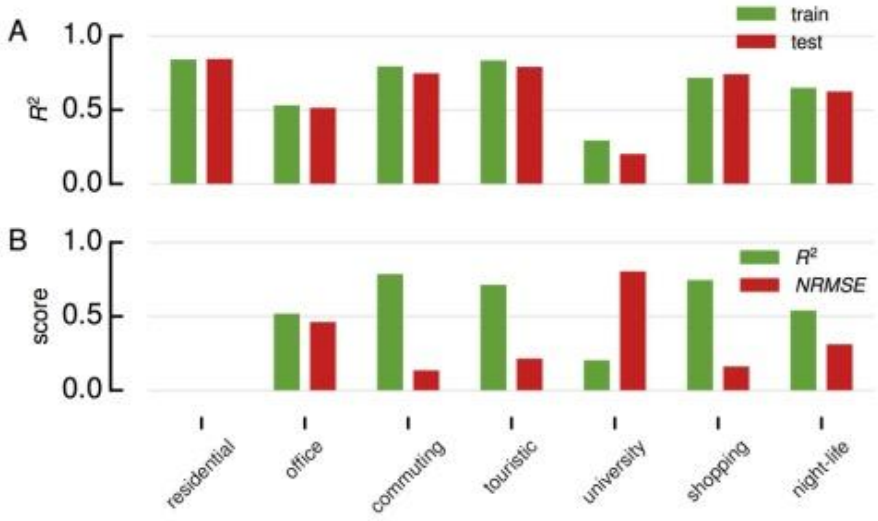

Figure 7Milan. Évaluation du modèle. A) R^2 pour les données d'apprentissage et de test, séparés par le type de land-use. B) R^2 et NRMSE du modèle appris sur des land-use résidentiels appliqué sur les autres land-use.

2.4 D'autre cas d'études

Nous généralisons notre analyse en considérant deux autres grandes villes en Italie, Rome et Turin. Pour chaque étude de cas en milieu urbain, nous adoptons la procédure de validation croisée, séparant les données d'apprentissage et de test. Nous estimons alors les paramètres du modèle $\hat{\alpha}$ et $\hat{\beta}$. La partie résidentielle (cf. Table. I) montre les résultats que nous obtenons, en termes de paramétrisation de $\hat{\alpha}$ et de$\hat{\beta}$. La partie droite de Table.1, notée mixte, se réfère à la qualité de l'estimation de toutes les zones, y compris celles qui ne sont pas de nature résidentielle: elle nous donne ainsi des informations sur la précision d'un modèle appris sur des données résidentielles seulement, lorsqu'il est utilisé sur une région urbaine complète . Dans ce second cas, nous utilisons les métriques de qualité moyennées suivantes:

$$\bar{R^2} = \sum_{\ell=1}^{L} \frac{N_\ell}{N} R_\ell^2, \qquad (6)$$

$$\overline{NRMSE} = \sum_{\ell=1}^{L} \frac{n_\ell}{N} NRMSE_\ell, \qquad (7)$$

où *L* est le nombre de différents land-use, *N*signifie le nombre de cellules spatiales associées à un land-use donné *l*, et R^2 (respectivement *NRMSE*) est le coefficient de détermination (respectivement, l'erreur quadratique moyenne normalisée) calculé sur leland-use *l*. Ainsi, ces mesures fournissent une moyenne pondérée de la performance de l'estimation sur tous les land-use. Les résultats dans Tab. I sont assez proches pour toutes les villes. La ville de Rome montre les scores les plus élevés, avec R^2 = 0,87 et *NRMSE*= 0,035 pour lesland-use résidentiels, et R^2= 0,84 et *NRMSE* = 0,044 pour le cas général. La ville de Milan suit Romede près, et Turin est légèrement en dessous. Néanmoins, le R^2que nous mesurons est toujours supérieur à celui observé dans les travaux précédents, par exemple [8]. De Tab. I, nous observons que les paramètres α et β ne sont pas significativement différents dans les trois villes. Nous explorons donc la possibilité d'estimer la population dans une zone urbaine en utilisant un modèle appris sur les données recueillies dans une autre ville. Fig. 8 résume nos conclusions. Dans l'ensemble, nous trouvons unR^2 élevé et une NRMSE faible dans tous les cas,ce qui nous permet de conclure qu'une estimation de la population des villesde façon croisée est possible. Cette observation est importante, ouvrant la voie à l'estimation des populations dans les villes pour lesquelles des données de trafic des réseaux mobilessont disponibles, mais où aucune vérite-terrain sur la distribution de la population est fourni.

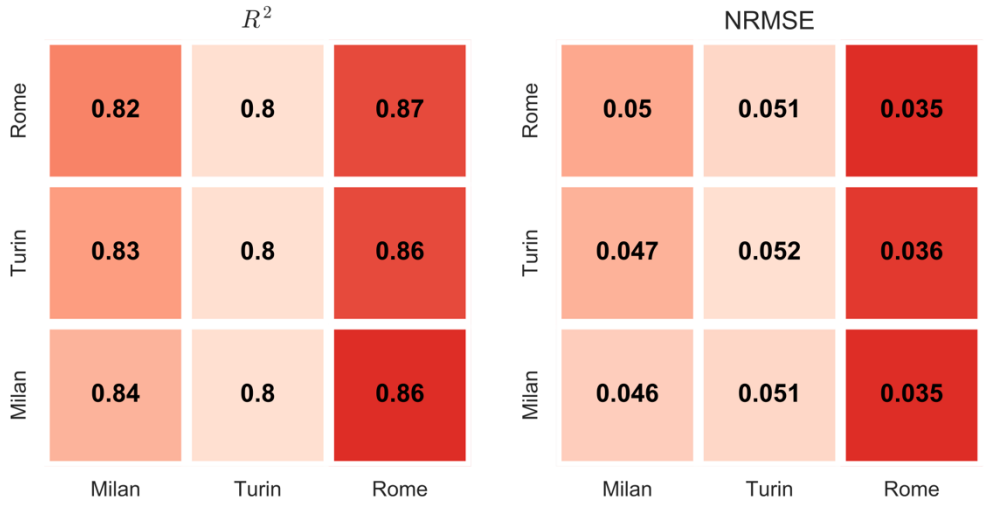

Figure 8Test de ville-croisée : les modèles appris sur les données recueillies dans les villes, le long des lignes est utilisé pour estimer la population des villes sur le long des colonnes. Les tableaux se réfèrent à R^2 et NRMSE.

	Residential								Mixed	
	Training						Test			
	$\hat{\alpha}$	$95\%C.I.$	$\hat{\beta}$	$95\%C.I.$	R^2	$NRMSE$	R^2	$NRMSE$	$\bar{R}^2$	$\overline{NRMSE}$
Milan	1.24	[1.03,1.37]	0.97	[0.89,1.0]	0.84	0.046	0.83	0.047	0.80	0.059
Turin	0.94	[0.82,1.2]	0.99	[0.88,1.2]	0.80	0.052	0.80	0.053	0.76	0.065
Rome	0.75	[0.67,0.98]	1.03	[0.92,1.1]	0.87	0.035	0.87	0.035	0.84	0.044

Tableau 1 R^2etNRMSE pour différente villes d'Italie.

3 Estimation de dynamiques de la population

Le problème principal dans l'estimation de la dynamique de la population est le manque de données deterrain, ce qui rend impossible l'apprentissage tel que décrit précédemment(3). Notre approche consiste à estimer une nouvelle relation

multivariée entre la distribution de la population, la présence des abonnés et le niveau d'activité des communications mobilesdes abonnés.

3.1 Présence des abonnés et niveau d'activités

Nous commençons par discuter de l'interaction entre la présence et le niveau d'activité des communications mobiles. Ce dernier est formellement défini comme étant la fréquence à laquelle un abonné interagit avec le réseau de téléphonie mobile.Fig. 9 représente le niveau moyen d'activité par abonné. On remarque une variation significative de l'activité, avec une décroissance de l'utilisation du réseau pendant la nuit et une augmentation des communications mobiles durant les heures de travail. Les différences entre les land-use sont modérées, comme indiqué sur la Fig. 10. Nous concluons que l'activité des communications mobiles est hétérogène, et un tel comportement émerge plutôt dans le temps que parland-use.Une telle hétérogénéité du niveau d'activité a un impact sur l'exactitude des informations de présence. Considérons à nouveau la Fig. 1: plus un dispositif mobile envoie ou reçoit des appels, des SMS, et des paquets de données, plus sa localisation est précise dans le jeu de données de présence. Une question légitime est alors si l'activité hétérogène que nous avons discutée peut être liée à la paramétrisation du modèle, et peut expliquer - en partie ou en totalité - la diversité des valeurs de$\hat{\alpha}$et$\hat{\beta}$observée dans Sec. 5.

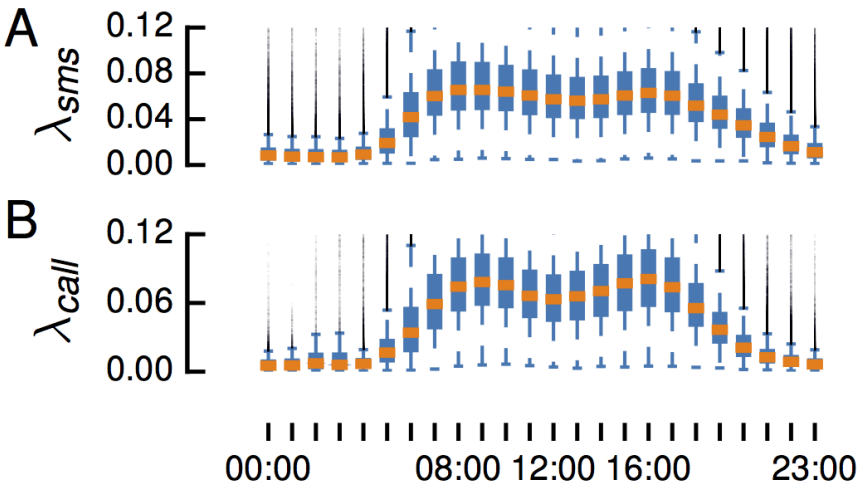

Figure 9le niveau d'activité des abonnés pour les appels vocaux (A, en haut) et: SMS (B, en bas) sur la journée.

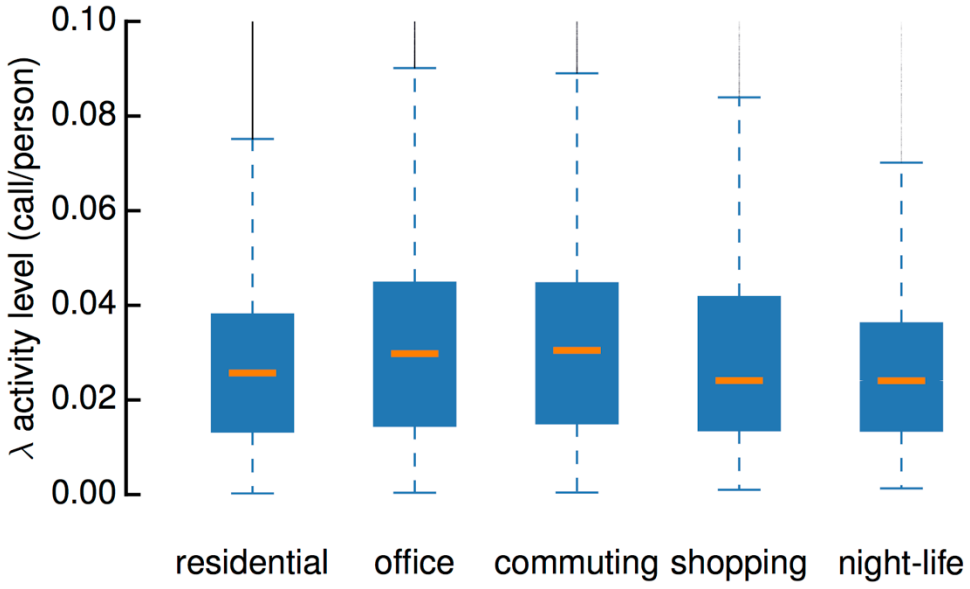

Figure 10le niveau d'activité des abonnés pour les appels, divisé par land-use

3.2 Estimation de la population avec le niveau d'activités

Nous étudions donc l'existence d'un lien entre le niveau d'activité des abonnés et les valeurs de α et β dans (2) qui régissent la relation entre la présence et la population. Nous ne disposons pas d'accès aux valeurs réelles, mais à leurs estimations $\hat{\alpha}$et$\hat{\beta}$. Nous recueillons ainsi des données dans toutes les villes qui se réfèrent à la période de la nuit, à savoir, de minuit à 8 heures : dans cette période les données du recensement ISTAT peuvent encore être considérées comme une vérité-terrain, puisque la plupart des gens seront à la maison. Nous traçons ensuite un diagramme du niveau d'activité λ en fonction des paramètres de régression$\hat{\alpha}$ et $\hat{\beta}$obtenus dans ces scénarios de (3). Les résultats sont représentés sur la Fig. 11.

Nous trouvons une relation linéaire importante entre λ et les deux paramètres. Les coefficients des modèles linéaires sont indiqués dans les plots. Ce résultat permet de dessiner un modèlemultivarié universel qui relie la densité de population à la fois à la présence des abonnés et au niveau d'activité des abonnés. Nous pouvons alors affiner notre modèle d'estimation :

$$\hat{\rho}_i(\lambda_i, \sigma_i) = (\hat{a}_\alpha \lambda_i + \hat{b}_\alpha) \cdot \sigma_i^{(\hat{a}_\beta \lambda_i + \hat{b}_\beta)}. \quad (8)$$

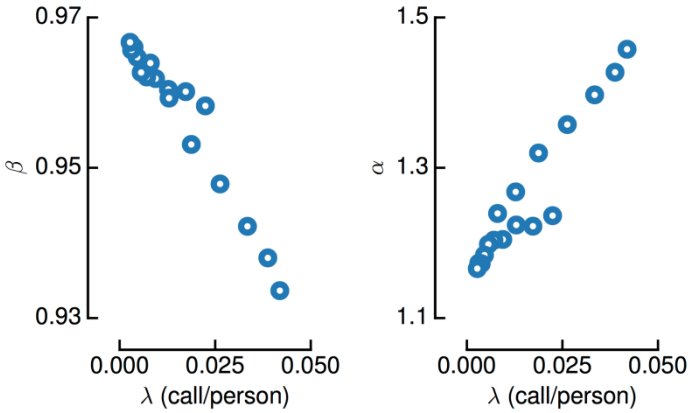

Figure 11Relation linéaire entre le niveau d'activité λ et le modèle des paramètres α̂ et β̂.

Une considération importante est que les nouveaux paramètres$a_\alpha b_\alpha, a_\beta, b_\beta$sont valables pour tous les scénarios, et sont compatibles avec les différents moments de la journée. Nous considérons donc que le modèledans (8) peut être utilisé de manière fiable pour l'estimation des populations dynamiques, étant donné que les séries temporelles de la présence des abonnés$\sigma_i$et du niveau d'activité des abonnés λ_i sont disponibles à partir des données du trafic des réseaux mobiles.

3.3 Un cas d'étude à Milan

La figure 12 illustre la répartition de la dynamique de la population à Milan et sa banlieue, déduite de notre modèle, à trois instants :le 15 avril 2015, à midi (A), le 22 avril 2015, à 17h (B) et le 19 avril 2015, à 10h (C). Dans chaque plot, les couleurs indiquent la variation de la population pendant l'heure correspondante :des flux élevés des personnes entrant à la cellule (rouge) aux flux élevés des personnes qui quittent la cellule (bleu). Il existe également des cellules neutres où la densité de la population ne varie pas au cours de la période considérée (blanc).Nous observons à 17 h, qu'il y aun flux important vers le centre-ville ou les zones commerciales (B, figure de droite). Le 19 avril est un samedi, et notre estimation de la population capte les mouvements des personnes vers des zones résidentielles en dehors de Milan (C, figure de gauche), ainsi que vers les zones de vie nocturne dans la ville (C, figure de droite).

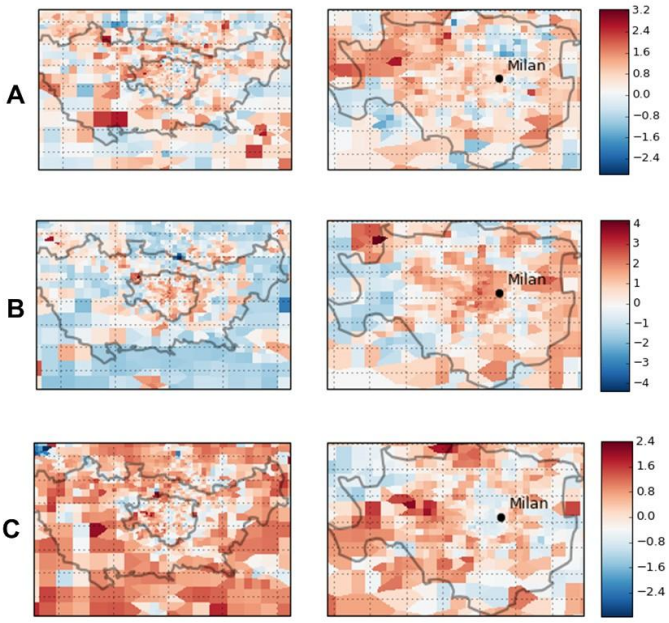

Figure 12Répartition de la dynamique de la population de Milan. A) 15Avril à midi. B) 22Avril à 17h. C) 19Avril 19 à 22h. Les Figuresde gauche montrent l'ensemble de la zone urbaine de Milan, celles de droite le centre ville uniquement.

4 Conclusions

Nous avons introduit une nouvelle approche [20] pour l'estimation de la population sur la base de la relation existant entre le volume d'activité mobile et la densité de population. Notre solution est fondée exclusivement sur les métadonnées collectées par les opérateurs des réseaux mobiles, à savoir sur la présence et sur le niveau d'activité des abonnées. Nos résultats démontrent commentnotre modèle permet une représentation statique fiable des populations dans différentes villes. Il surpasse également les propositions antérieures de la littérature, grâce à l'utilisation de métadonnées plus appropriéeset au filtrage temporel. Finalement, nous avons appliqué notre modèle pour l'estimation de la distribution dynamique des populations à partir de leurs activités quotidiennes.Les tests de preuve de concept ont montré une cohérence prometteuse.

4.1 Documents de référence

[1] C. Song,Z. Qu,N.Blumm,andA.-L.Barabási,“Limitsof predictability inhumanmobility,”Science, vol.327,no.5968,pp.1018–1021,2010.

[2] J.P.Bagrow,D.Wang,andA.-L.Barabasi,“Collectiveresponseof humanpopulationstolarge-scaleemergencies,”PloSone,vol.6,no.3, p.e17680,2011.

[3] P. Bajardi, C. Poletto, J. J. Ramasco, M. Tizzoni, V. Colizza, and A.Vespignani, “Humanmobilitynetworks,travelrestrictions, andthe globalspreadof 2009h1n1pandemic,”PloSone,vol.6, no.1, p. e16591,2011.

[4]Y.Yang,C.Herrera,N.Eagle,andM.C.González, “Limitsof predictabilityincommutingflowsinthe absenceofdata forcalibration,” NatureScientificReports,vol.4,no.5662,2014.

[5]N.Caceres,L.Romero,F.Benitez,andJ.Castillo,“Traffic flow estimationmodelsusingcellularphonedata,”IEEETransactions on IntelligentTransportationSystems,vol.13,no.3,2012.

[6] D.Naboulsi,M.Fiore,R.Stanica,andS.Ribot,“Large-scale mobile trafficanalysis: asurvey,”IEEECommunicationsSurveysandTutorials, vol.18,no.1,2016.

[7] G.Krings,F.Calabrese,C.Ratti,andV.D.Blondel,“Urbangravity: amodelforinter-citytelecommunicationflows,”Journal ofStatistical Mechanics:TheoryandExperiment,vol.2009, no.07,p.L07003,2009.

[8] R.W.Douglass,D.A.Meyer,M.Ram,D.Rideout,andD.Song, “Highresolutionpopulationestimatesfromtelecommunications data,” EPJDataScience,vol.4,no.1,pp.1–13,2015.

[9] P. Deville, C. Linard, S. Martin, M. Gilbert, F. R. Stevens, A. E.Gaughan,V.D.Blondel,andA.J.Tatem,“Dynamicpopulationmapping usingmobilephonedata,”Proceedings ofthe National Academyof Sciences,vol.111,no.45,pp.15888–15893,2014.

[10] B. C. Csáji, A. Browet, V. A. Traag, J.-C. Delvenne, E. Huens, P.VanDooren,Z.Smoreda,andV.D.Blondel,“Exploring themobility ofmobilephoneusers,”PhysicaA: StatisticalMechanicsandits Applications,vol.392,no.6,pp.1459–1473,2013.

[11] S.Bekhor,Y.Cohen,andC.Solomon,“Evaluatinglong-distance travel patternsinisraelbytrackingcellularphonepositions,” Journalof AdvancedTransportation,vol.47,no.4,pp.435–446,2013.

[12] F.Calabrese,G.DiLorenzo,L.Liu,andC.Ratti,“Estimating origin- destinationflows usingmobilephonelocationdata,”IEEEPervasive Computing,vol.10,no.4,pp.0036–44,2011.

[13] C.Ratti,R.Pulselli,S.Williams,andD.Frenchman,"Mobileland- scapes:Usinglocationdatafromcell-phones forurbananalysis,"Environmentand PlanningBPlanningandDesign,vol.33,no.5,2006.

[14] C.Kang,Y.Liu,X.Ma,andL.Wu,"Towards estimating urbanpopula- tiondistributionsfrommobilecalldata,"Journalof UrbanTechnology, vol.19,no.4,pp.3–21,2012.

[15]Telecomitaliabigdatachallenge. [Online].Available: http://www. telecomitalia.com/tit/en/innovazione/big-data-challenge-2015.html.

[16] Istat.[Online]. Available: http://www.istat.it/en/population-and-housing-census

[17] A.Furno,R.Stanica,andM.Fiore,"Acomparative evaluationofurban fabric detectiontechniquesbased onmobile trafficdata," inProceedings ofthe2015IEEE/ACMInternationalConferenceonAdvancesinSocial NetworksAnalysisandMining2015. ACM,2015,pp.689–696.

[18] B.Cici,M.Gjoka,A.Markopoulou, andC.T.Butts,"Onthedecom- positionofcellphoneactivitypatternsandtheirconnection withurban ecology,"inProceedingsofthe16thACMInternationalSymposiumon MobileAdHocNetworkingandComputing. ACM,2015, pp.317–326.

[19] M.A.FischlerandR.C.Bolles,"Randomsampleconsensus:aparadigm formodelfitting withapplicationstoimageanalysisandautomated cartography,"CommunicationsoftheACM,vol.24,no.6,pp.381–395,1981.

[20] Ghazaleh Khodabandelou, Vincent Gauthier, Mounim El-Yacoubi, Marco Fiore, « Population Estimation from Mobile Network Traffic Metadata », To Appear in Wowmom, June, 2016.