Université Gustave Eiffel

Multi-source data fusion for the analysis of mobility

Thesis for the degree of
Doctor of Université Gustave Eiffel

by

Benoît Matet

presented on 23rd April 2024 before:

Latifa Oukhellou Professor, Université Gustave Eiffel	(Director)
Nour-Eddin El Faouzi Professor, Université Gustave Eiffel	(Director)
Francesco Viti Associate Professor, Université du Luxembourg	(Reviewer)
Lijun Sun Associate Professor, McGill university	(Reviewer)
Karine Bennis Zeitouni Professor, Université Paris-Saclay	(Examiner)
Eleni Vlahogianni Professor, National Technical University of Athens	(Examiner)
Étienne Côme Researcher, Université Gustave Eiffel	(Examiner)
Angelo Furno Researcher, ENTPE	(Examiner)

Remerciements (Acknowledgements)

J'ai pu bénéficier de l'aide et du soutien de nombreuses personnes lors de la réalisation de cette thèse, que je tiens à remercier ici. Tout d'abord Latifa Oukhellou, qui a porté cette thèse en particulier mais qui de façon plus générale a toujours à cœur le bien-être de ses doctorants et qui les protège des aléas administratifs avec brio. Ensuite mes autres encadrants, Étienne Côme, Angelo Furno et Nour-Eddin el Faouzi, qui avec Latifa ont rendu cette thèse possible par leur encadrement et leurs conseils. Je tiens à remercier également Marco Fiore et Sebastian Hörl, avec qui j'ai eu la chance de travailler et qui m'ont fait bénéficier de leurs retours positifs et constructifs.

I am grateful to Francesco Viti and Lijun Sun for having accepted to be the reviewers of this work and for their insightful feedbacks. These thanks also go to Eleni Vlahogianni and Karin Bennis Zeitouni for participating to this jury as examiners.

Cette thèse n'aurait même pas commencé sans l'assistance avisée de Gabriel Jouffrai, mon excellent maître de stage qui m'a très aimablement recommandé au projet. Ces remerciements vont aussi à Andreea Lachapelle, dont l'intelligence pédagogique m'a montré l'importance et les joies de la recherche.

Je remercie également mes collègues, doctorants et docteurs, avec qui j'ai pu partager des séminaires enrichissants, des TDs matinaux, ou des commentaires désobligeants sur Latex : Paul, Thomas, Louise, Mostafa, Negin, Khadidja, Rodolphe, Paul, Hugues...

Je me dois bien sûr de remercier mes amis de collège et de lycée, qui par leurs singeries sans cesse renouvelées font de chaque jour une nouvelle surprise. Je veux parler d'Édouard, François, Mathilde, Maxime et Maxime, Olivier, Baptiste, Arthur et Mélanie, et bien évidemment l'inégalable Julien.

Cette liste ne serait pas complète sans remercier mes amis d'école, avec qui j'ai pu découvrir à quoi ressemblait le monde en dehors de la prépa : Lucie, Jihane-Louise, Sylvain, Selim et Anna, Yacine, Rachel, Victor, Achraf et Thomas.

Enfin, je me dois de remercier ma famille, qui a toujours su me guider lorsque j'en avais besoin. Tout particulièrement ma tante Manou pour ses bons conseils en maths et en lectures, et mes frères et sœur Caroline, Nicolas et Simon, sur qui j'ai pu compter dans toutes les épreuves rencontrées. Enfin, je tiens à remercier mes parents et mes grands-parents, qui m'ont inculqué la curiosité, la rigueur, et la patience nécessaires à mener cette thèse à bien.

Abstract

This thesis focuses on generating and using safe and accurate data to describe the mobility of people in an urban environment. The main focus of this work is the Origin-Destination (OD) matrices obtained from mobile phone data, which describe the flows of population between the zones of a city. This data is characterized by huge volumes, which call for light-weight processing solutions, and a high variety, which imply a privacy risk for outliers.

In a first part, we develop an algorithm to efficiently guarantee the k-anonymization of such OD matrices *via* generalization and suppression. Our method implements a hard constraint on the number of trips that can be suppressed, in order to maintain the representativity of the data. The spatial generalization is formalized as a knapsack problem with a dependency tree, whose dual can be efficiently solved using the Some Breakpoints Algorithm. We also study the relaxed problem, which does not guarantee a maximum number of suppressed trips but instead a maximum level of generalization. We compare our approaches to an extensive benchmark of the state of the art in anonymization on a collection of large-scale OD matrices.

In a second part, we propose two steps to generate more realistic synthetic travel demand, derived from a Household Travel Survey (HTS), using dynamic OD matrices. In the first step, we calibrate the temporal distribution of trips made during the day by formalizing it as hierarchical population problem. In the second step, we draw activity locations using the OD matrices as transition probabilities in a probabilistic graph model. We illustrate a pitfall in the estimation of such a model when implementing basic agenda constraints, such as the fact that the "home" locations must all be equal. These added dependencies create cycles in the graph which invalidate the direct use of the OD matrices as maximum likelihood estimators. Instead, we propose an heuristic adaptation to estimate the parameters of the model. Then, we implement a variety of approaches corresponding to different trade-offs between matching the OD matrices and matching the surveys. This allows us to give a quantitative measure of the discrepancies between the OD matrices and the HTS, which are known to exist but hard to measure as the two sources do not describe the same objects.

This work takes place in a context of a recent multiplication of available data sources for transportation studies. In particular, passive sources such as mobile phone data collect traveler's information without input on their part and mostly without their knowing. They carry invaluable insights into the dynamics of travel demand due to their unequaled penetration rate, but are also an ethic liability due to their monitoring potential. By guaranteeing a foolproof anonymization of the data and illustrating its use in travel demand synthesis, we aim at addressing its problem of privacy while at the same time leveraging it to produce a realistic, exhaustive overview of urban transportation.

Keywords: Mobility, OD matrices, Anonymization, Generalization and suppression, Synthetic travel demand, Probabilistic graph models.

Résumé

Cette thèse explore la génération et l'utilisation de données sûres et précises pour décrire la mobilité des personnes en environnement urbain. Une attention particulière est donnée aux matrices Origine-Destination (OD) obtenues à partir des données de téléphonie mobile, qui décrivent les flux de population entre les zones d'une ville. Ces données sont caractérisées par de gros volumes, qui nécessitent des solutions de traitement légères, et une grande variété, impliquant un risque pour la vie privée des personnes effectuant des déplacements peu communs.

Dans une première partie, nous développons un algorithme pour garantir efficacement la k-anonymisation de telles matrices OD par généralisation et suppression. Notre méthode implémente une contrainte dure sur le nombre de déplacements pouvant être supprimés, afin de maintenir la représentativité des données. La généralisation spatiale est formalisée comme un problème de sac à dos avec arbre de dépendances, dont le dual peut être résolu efficacement à l'aide du "Some Breakpoints Algorithm". Nous étudions également les propriétés de la relaxation du problème, qui ne garantit pas un nombre maximum de déplacements supprimés mais plutôt un niveau maximum de généralisation. Nous comparons nos approches à une variété de méthodes d'anonymisation de l'état de l'art sur une collection de matrices OD à grande échelle.

Dans une deuxième partie, nous proposons deux étapes pour générer une demande synthétique de déplacements, basée sur une enquête de transport, plus réalistes à l'aide de matrices OD dynamiques. Dans un premier temps, nous calibrons la répartition temporelle des déplacements effectués dans la journée en la formalisant comme un problème de population hiérarchique. Puis nous tirons les emplacements d'activité en utilisant les matrices OD comme probabilités de transition dans un modèle graphique probabiliste. Nous illustrons un écueil dans l'estimation d'un tel modèle lors de la mise en œuvre de contraintes d'agendas, telles que le fait que toutes les activité "Domicile" doivent avoir lieu au même endroit. Ces contraintes créent des cycles dans les graphes, qui invalident l'utilisation directe des matrices OD comme estimateurs du maximum de vraisemblance. Nous remplaçons cet estimateur par une adaptation heuristique, et nous proposons plusieurs structures de graphes correspondant à différents compromis entre le respect des matrices OD et des enquêtes. Cela permet de donner une mesure quantitative des écarts entre ces deux sources, dont l'existence est connue mais difficile à mesurer comme les deux sources ne décrivent pas les mêmes objets.

Ce travail s'inscrit dans le contexte de l'apparition de sources passives dans les études de transports, qui collectent des informations sur les voyageurs sans intervention de leur part et généralement à leur insu. Elles apportent des informations précieuses sur la dynamique des trajets en raison de leur taux de pénétration inégalé, mais constituent également une responsabilité éthique en raison de leur potentiel de contrôle sur la population. En garantissant une anonymisation à toute épreuve des données et en illustrant leur utilisation dans la synthèse de demande de déplacement, nous visons à répondre au problème de la vie privée tout en les exploitant pour produire un aperçu réaliste et exhaustif des transports urbains.

Mots-clefs: Mobilité, matrices OD, Anonymisation, Generalisation et suppression, Demande en transport synthétique, Models graphiques probabilistes.

Résumé long

Le début du XXIème siècle marque le début de la période où plus de la moitié de la population mondiale vit dans des zones urbaines (United Nations, 2018). Ce nombre devrait atteindre les deux tiers d'ici 2050, avec environ un tiers des citadins vivant dans des villes de plus d'un million d'habitants. La bonne organisation de la mobilité au sein de ces villes est un facteur majeur de la qualité de vie globale de leurs habitants, des pollutions atmosphérique et sonore auxquelles ils sont exposés, et de l'accessibilité des différents points d'intérêts. Les réseaux de transports urbains inappropriés présentent un coût social, mais aussi plus pragramtiquement des coûts économiques : En Europe, on estime actuellement que la congestion urbaine coûte au total 110 milliards par an (Brannigan et al., 2017).

Afin de concevoir un réseau de transport adapté, il est nécessaire de connaître les besoins de ses utilisateurs. L'étude du comportement de mobilité des personnes, quel que soit l'aspect spécifique mesuré, est regroupée sous le terme générique **demande de transport**. Certaines applications à but lucratif des études en demande de transport consistent à identifier les meilleurs emplacements où afficher des publicités, ou à décider des nouveaux sites pour une chaîne de magasins. D'autres applications, plus orientées vers le service publique, visent à suivre et comprendre l'impact des politiques publiques. Une compréhension plus approfondie de la dynamique du système peut également aider à identifier comment créer des comportements souhaitables dans les réseaux de transport : par exemple, minimiser le temps total passé dans les embouteillages, minimiser l'empreinte carbone des déplacements, maximiser l'utilisation du vélo, inciter les automobilistes à se tourner vers les transports en commun, ou tout autre critère jugé bénéfique pour le territoire ou sa population.

Cette thèse est motivée par la nécessité de concevoir de nouveaux réseaux de transport durables dans les villes modernes, qui remplacent la prépondérance de la voiture par un réseau de transports publics plus résilient et plus efficace. Les infrastructures de transport, notamment ferroviaires, sont remarquablement coûteuses : les tramways français coûtent en moyenne 34 millions d'euros par kilomètre (Bensalah et al., 2017), tandis que les coûts d'excavation font monter encore plus les prix des lignes de métro. À titre de comparaison, exprimé en dollars américains de 2015, l'extension de la ligne Jubilee de 1999 à Londres a coûté 505 millions de dollars par kilomètre, tandis que l'extension de la ligne IRT Flushing de 2015 à New York, une ville connue pour ses coûts prohibitifs de construction de métro, a été estimée à 1,07 milliard de dollars par kilomètre (Hess, 2016). Les 205 kilomètres du projet Grand Paris Express, le plus grand projet de transport urbain en construction au monde, devraient coûter à terme environ 42 milliards d'euros, pour une moyenne en dollars américains d'environ 220 millions de dollars par kilomètre (Commission des finances, 2020).

Il est donc fondamental de pouvoir comprendre en détail le comportement de mobilité de la population concernée, afin que les ressources investies dans les infrastructures soient utilisées de la manière la plus efficace possible. Cependant, à ce jour ces connaissances se limitent à des indicateurs agrégés et statiques, et aucun responsable ou expert ne peut, par exemple, donner une estimation du nombre de personnes qui se sont déplacées d'un point de la ville à un autre, hier à 10 heures. En fait, la compréhension de la mobilité urbaine est si limitée que certains résultats contre-intuitifs restent encore à expliquer. Dans un nombre croissant de cas documentés, il a été observé que la fermeture d'une route ne génère pas plus de trafic dans la

zone voisine (Cairns et al., 2002). Ce que deviennent les usagers de la route manquants reste encore flou, au point que l'on parle d'« évaporation du trafic ».

Il y a deux raisons pour lesquelles il est difficile d'estimer exactement où se trouve et où va une population. Premièrement, comme les villes sont des systèmes complexes dans lesquels les personnes, les organisations et les véhicules interagissent de manières diverses et complexes, la modélisation d'un système de transport nécessite de modéliser les décisions de ses utilisateurs, leurs préférences et les influences qu'ils exercent les uns sur les autres. Comme le dit le physicien Murray Gell-Mann : « Imaginez à quel point la physique serait difficile si les particules réfléchissaient. » Deuxièmement, détailler les déplacements de personnes est généralement mal vu dans un monde où la vie privée est considérée comme un droit fondamental. S'il avait été plus sensible à ces préoccupations, Murray Gell-Mann aurait pu ajouter: « Et comme ça serait difficile si les particules ne voulaient pas que vous les mesuriez. »

Plusieurs sources de données sont disponibles pour mesurer une partie de ce comportement de mobilité. Elles sont toutes incomplètes à leur manière et mesurent généralement des concepts différents dans des contextes différents, ce qui rend difficile leur utilisation conjointe. Fusionner ces sources de données pour comprendre les détails du trafic urbain, tout en garantissant la confidentialité des sources disponibles, a été le fil directeur des travaux présentés ici.

Le moyen le plus évident de mesurer la mobilité d'une grande partie de la population serait d'accéder aux GPS des téléphones portables. Cette première possibilité existe, et donne ce que l'on appelle des données de Location Based Services (LBS). Les données LBS sont émises par des applications mobiles (les services) qui sondent la localisation de l'appareil, généralement pour des raisons opérationnelles, et la transmettent au collecteur de données pour un traitement supplémentaire. Elles peuvent donner des résultats intéressants, mais elles ne concernent souvent qu'un nombre limité de personnes, souvent les utilisateurs d'une application spécifique. La deuxième source la plus évidente serait les tickets de transports en commun : dans la plupart des systèmes de métro modernes, il est obligatoire de valider un pass personnel à l'entrée et à la sortie du métro, permettant de collecter ce qu'on appelle les **matrices Origine-Destination (OD)** des voyageurs. Cependant, les données de billetique ne peuvent mesurer que l'utilisation des transports publics. De plus, dans le cas des bus ou des tramways où il n'est en général pas nécessaire de valider à la sortie, on n'obtient qu'une carte des montées des voyageurs au lieu d'une matrice OD.

Antérieures à ces sources modernes, les études sur la demande de voyages reposaient simplement sur des enquêtes, demandant aux répondants quelques informations personnelles et surtout un rapport détaillé sur un jour de leur semaine. De telles enquêtes axées sur les transports sont appelées **Enquêtes sur les transports des ménages (HTS)**. Elles sont précieuses tant par leur niveau de détail que par leur gestion de la cohérence individuelle : plutôt que de simples flux décrits par des matrices OD, les HTS décrivent l'agenda complet des individus pour une journée donnée, avec le lieu, le but et les heures de chaque activité, mais également le mode de transport utilisé entre les activités. Cela permet de modéliser les interactions entre les activités et entre les individus, ce qui constitue la pierre angulaire des modèles de mobilité modernes. En fait, les modèles de mobilité sont conçus pour prendre en entrée des données qui ressemblent à un HTS mais décrivant une population complète.

De telles données doivent être synthétisées, d'où le problème de créer une **demande de déplacement synthétique**. Il est très difficile de créer une demande synthétique complète de voyages uniquement à partir d'un HTS, car ces derniers sont limités par leur très faible nombre de répondants, aggravé par leur taux élevé de non-réponse. Cela les rend peu fiables pour l'estimation des localisations, dans la mesure où seule une poignée d'emplacements dans la ville seront effectivement visités par l'un des répondants du sondage. D'un point de vue statistique, si la ville est divisée en 200 zones (ce qui est une division relativement grossière), il y a alors 320 milliards de possibilités pour une chaîne de 5 activités, et aucune enquête ne peut donner une estimation fiable de la probabilité de chacune de ces possibilités.

Une récente source possible pour mesurer les mouvements de population est les **données de téléphonie mobile**. Contrairement aux traces GPS capturées par les applications mobiles, ce que nous appelons données de téléphonie mobile sont les données opérationnelles collectées par les opérateurs mobiles. Cela inclut deux sources : les **Call Detail Records (CDRs)**, qui sont collectés à des fins de facturation, et **Network Signalization Data (NSD)**, qui correspondent à tous les messages échangés dans le cadre du fonctionnement normal du réseau. Les deux sources enregistrent, entre autres, un horodatage, l'identifiant d'une antenne de téléphone portable et l'identifiant de la carte SIM d'un utilisateur mobile, indiquant que l'utilisateur se trouvait probablement à proximité de l'antenne à l'heure indiquée. Le principal avantage des données de téléphonie mobile est leur large emprise : presque tout le monde possède un téléphone et le marché est divisé parmi seulement une poignée d'opérateurs, ce qui signifie qu'un opérateur peut facilement recueillir des informations sur une portion significative de la population. À ce titre, les données mobiles constituent un atout prometteur vers une compréhension exhaustive de la mobilité urbaine.

Leur utilisation fait naître l'espoir de pouvoir un jour comprendre la demande de déplacement avec des approches rapides et peu coûteuses, capables de produire des informations sur la dynamique des transports dans un délai très court et, à terme, d'anticiper la demande de déplacement dans des situations hypothétiques. Cependant, avant que ces espoirs puissent se réaliser, deux problèmes principaux doivent être résolus :

- Le premier problème est celui de la précision et de la correction. La précision des données mobiles se limite à l'indication d'une antenne de téléphone portable, dont la zone de couverture réelle est généralement inconnue, peu pratique à utiliser et grande par rapport à la distance séparant les différents points d'intérêt d'une ville. La précision temporelle des CDR est largement admise comme insatisfaisante en raison du faible taux d'échantillonnage, un problème atténué mais pas entièrement résolu par les NSD (Bonnetain, 2022). De plus, il n'y a pas de détail autre que des points spatio-temporels, et on pourrait dire que la précision en termes de mode de transport ou de profils socio-économiques est nulle. Les données mobiles peuvent également être incorrectes, dans le sens où les informations, même précises, sont factuellement fausses. En effet, la logique qui régit la connexion des téléphones mobiles aux antennes du réseau dépend de facteurs multiples et incontrôlés, et les spécifications techniques des communications par antenne téléphonique visent à maximiser la qualité de service, mais pas la facilité d'utilisation pour les études de mobilité. Il en résulte des trajectoires bruitées et chaotiques avec des déplacements manquants ou à l'inverse des artefacts. Interpréter directement les données mobiles comme des déplacements de population mène aussi à des erreurs d'estimation

dues au fait que la clientèle de chaque opérateur mobile n'est pas parfaitement représentative de la population dans son ensemble.

- Le deuxième problème est bien sûr celui de la vie privée, car malgré les problèmes évoqués ci-dessus, les données mobiles sont avant tout un outil prometteur pour suivre les déplacements d'une population à une échelle inédite sans aucune intervention des individus ciblés. Il est tout à fait naturel de s'attendre à ce qu'une source de données aussi riche soit totalement anonyme, au point de pouvoir être rendue publique. Après tout, ce sont des informations sur la vie publique, utilisées pour contribuer à l'élaboration de politiques publiques. Cette responsabilité en matière de vie privée constitue également un obstacle à la diffusion des données. Dans les pays européens, le RGPD (European Commission, 2018) particulièrement restrictif dissuade fortement les propriétaires de données de mobilité de partager toute forme de données, y compris de simples matrices OD. Cela entrave les efforts de recherche visant à développer de meilleures méthodes de planification des transports pour des réseaux de transport plus écologiques, plus sûrs et plus résilients. On peut noter que ce problème n'est pas une fatalité : le RGPD vise à protéger les informations personnelles, alors que les études sur les transports s'intéressent normalement davantage aux tendances générales de la mobilité.

L'objectif principal de ce travail est de promouvoir des études de mobilité sûre utilisant les données mobiles. Pour cela, nous étudions particulièrement l'utilisation de matrices OD, car elles sont par nature plus susceptibles d'être totalement anonymisées que les trajectoires complètes. Nous considérons que les matrices OD dérivées des données de téléphonie mobile constituent une bonne opportunité pour introduire une approche basée sur les données pour la spatialisation de la demande de déplacement synthétique. Les données mobiles fournissent des informations spatio-temporelles détaillées, dynamiques sur la localisation de la population sans détails individuels, tandis que les HTS fournissent des descriptions individuelles riches qui manquent de représentativité spatiale. La combinaison des deux sources constitue une opportunité de synthétiser une population globale qui soit réaliste au niveau individuel tout en se comportant de manière réaliste au niveau de la population.

La responsabilité en matière de confidentialité d'un ensemble de données se mesure comme sa vulnérabilité aux trois types d'attaques (Hörl and Axhausen, 2021) :

- Attaques de liaison d'enregistrement, lorsqu'un attaquant tente d'identifier sa cible dans l'ensemble de données ;
- Attaque de liaison d'attributs, lorsqu'un attaquant déduit la valeur d'une variable particulière de sa cible ;
- Attaques probabilistes, lorsque l'attaquant met à jour ses informations préalables de quelque manière que ce soit sur la cible.

Chacune de ces trois attaques est un assouplissement de la précédente, où l'objectif est moins ambitieux pour l'attaquant mais le succès est plus probable.

Le critère d'anonymisation le plus répandu est la ***k*-anonymisation** (Sweeney, 2002a). Il s'agit d'une garantie formelle de confidentialité utilisée à la fois dans la recherche et par des autorités telles que le régulateur français CNIL. Un ensemble de données est dit k-anonyme si

chaque individu contenu dans les données est impossible à distinguer de $k-1$ autres individus. Techniquement, la k-anonymité à partir de $k \geq 2$ protège contre les attaques *record linkage*, bien que cela puisse être considéré comme une condition d'anonymat plutôt faible. En particulier, les faibles valeurs de k offrent une protection très limitée contre les *attribute linkage* et les attaques probabilistes. L'état de l'art vise généralement des k compris entre 3 et 5 pour les ensembles de données difficiles à anonymiser, et jusqu'à 200 pour les ensembles de données simples (Liang and Samavi, 2020).

L'approche principale pour atteindre la k-anonymisation est la **généralisation et suppression** (Sweeney, 2002a), c'est-à-dire la réduction de la précision des données pour former des groupes indiscernables d'au moins k individus et la suppression des individus qui ne font pas partie de ces groupes. Dans le cas des matrices OD, cela signifie que l'origine et la destination des flux sont fusionnées entre elles jusqu'à ce que les sommes des volumes des flux atteignent plus que k. Trouver une telle solution qui minimise la perte de précision est connu pour être un problème NP-difficile (Bettini et al., 2005). Historiquement, le premier algorithme d'anonymisation k était Datafly (Sweeney, 2002a), qui s'appuie sur une hiérarchie de généralisation décrivant comment les modalités doivent être fusionnées. Si l'on choisit de généraliser l'ensemble des données à un niveau supérieur, alors les modalités possibles sont les parents des feuilles. Datafly trouve la meilleure coupe horizontale dans la hiérarchie, ce qui signifie que toutes les données sont généralisées au même niveau. Cette approche de **généralisation uniforme** a l'avantage de pouvoir s'adapter à d'énormes volumes de données, ainsi qu'à des données possédant de nombreux attributs. Dans le but de trouver un résultat plus fin, certaines approches recherchent une généralisation au niveau individuel, ce qui donne une solution proche d'un *clustering* (Liang and Samavi, 2020). Dans le domaine spécifique des données de mobilité, cette approche est mieux représentée par Glove (Gramaglia and Fiore, 2015), qui généralise des points dans un ensemble de données de trajectoires. Cependant, ces approches passent mal à l'écjelle pour les données caractérisés par un grand nombre de flux, comme c'est le cas pour les données dérivées des CDR et NSD des téléphones mobiles : par exemple, Liang and Samavi (2020) rapporte des temps de calcul de l'ordre de quelques heures pour de faibles valeurs de k, et on peut estimer que leur approche prendrait des jours pour $k \geq 10$ étant donné sa dépendance temporelle exponentielle. Pour anonymiser des matrices OD de façon rapide et précise, un algorithme dédié est nécessaire.

Dans ce travail, nous développons une approche légère pour l'anonymisation k des matrices OD qui exploite la faible dimension des données pour explorer un espace de solutions plus grand que les algorithmes de généralisation classiques, tout en gardant les restrictions pertinentes de l'espace de recherche afin d'être évolutif sur des matrices à nombre de flux élevé. Nous l'appliquons à une variété de matrices OD réelles à grande échelle collectées par la New York City Taxi and Limousine Commission et dérivées du défi Data for Development (D4D) organisé par Orange au Sénégal et en Côte d'Ivoire. Comparée à un benchmark étendu composé d'algorithmes de généralisation réguliers ainsi que d'approches d'anonymisation de la mobilité, nous montrons que notre méthode est 27% plus précise et 9 fois plus rapide que des approches comparables capables de s'adapter aux mêmes ensembles de données.

Nous étudions aussi le problème de la synthèse de la demande en transport avec et sans l'utilisation de données mobiles. Nous nous concentrons sur le processus de l'état de l'art introduit par Hörl and Balac (2021), qui repose sur une succession d'étapes pour intégrer

diverses sources de données. Comme première contribution à cet égard, nous proposons une méthode de calibrage afin que la distribution temporelle des déplacements effectués pendant la journée corresponde à la distribution observée à partir des données mobiles. En formalisant ce problème comme un problème de **population hiérarchique**, nous pouvons ajuster le nombre de déplacements effectués par la population à chaque pas de temps de la journée tout en conservant la composition socio-économique de la population synthétique. Cette étape peut être appliquée *a posteriori* après l'attribution des chaînes d'activités, quelle que soit l'approche exacte utilisée pour créer la population synthétique. Nous trouvons une solution par l'algorithme **Iterative Proportional Updating (IPU)**, qui est une approche assez simple sans justification formelle de correction, mais nous observons qu'elle donne des résultats fiables sans avoir besoin de calibrer les paramètres. Au contraire, résoudre le problème par un algorithme de descente de gradient tel que Adam (Kingma and Ba, 2014) nécessite de fixer un coefficient de régularisation qui implique un compromis entre un bon ajustement des distributions marginales et une faible distorsion de la population utilisée en entrée.

Enfin, nous étudions l'utilisation des matrices OD dans le problème de spatialisation de la demande de transport. L'attribution de ces localisations reste un problème encore ouvert du fait à la fois du manque de données disponibles en la matière et du caractère très dimensionnel du problème : même dans un contexte simplifié d'une zone d'étude divisée en quelques centaines de quartiers, une chaîne de localisations peut contenir jusqu'à dix emplacements pour un total de 10^{100} de possibilités. Bien sûr, une structure solide existe dans cette distribution, et la plupart de ces possibilités de 10^{100} seraient physiquement impossibles ou du moins hautement improbables. Tout d'abord, on peut s'attendre à ce que chaque individu dispose d'un seul endroit pour toutes ses activités « à la maison », et dans le cas général, il en va de même pour toutes ses activités « de travail » et « d'études ». Cela conduit l'état de l'art à considérer ces **localisations primaires** séparément des autres, car elles sont considérées comme des points d'ancrage fixes indépendants du reste de l'agenda (Bowman and Ben-Akiva, 2001).

Les méthodes populaires pour tracer une localisation secondaire entre deux activités principales considèrent l'ensemble des points spatio-temporels pouvant être atteints entre la fin de l'activité précédente et le début de la suivante, étant donné une certaine vitesse maximale. Cet ensemble est un **prisme spatio-temporel** (Lenntorp, 1976). Dans la synthèse de la demande de transport, le prisme spatio-temporel agit comme une restriction de l'ensemble des emplacements possibles parmi lesquels choisir (Yoon et al., 2012; Justen et al., 2013). D'autres approches basées sur les données, telles que Anda et al. (2021), modélisent la distribution conjointe des emplacements, des horaires ou des modes de transport sous la forme d'un modèle graphique probabiliste, dont les paramètres sont ensuite estimés à l'aide de matrices OD et d'enquêtes.

Nous développons une étape de spatialisation comparable à celle propsoée par Anda et al. (2021), en modélisant les chaînes de localisation comme un modèle de graphe probabiliste dont les paramètres sont estimés à partir des matrices OD. Par rapport aux autres approches basées sur des graphes, notre travail intègre la spatialisation dans le processus plus général décrit par Hörl and Balac (2021). Nous discutons également d'une méthodologie pour estimer ces modèles, car leurs structures générales non arborescentes nous empêchent d'interpréter directement les matrices OD comme des probabilités de transition. Nous proposons une variété de structures de graphes qui offrent différents compromis entre la re-création des matrices OD

au niveau de la population, et le respect d'agendas réalistes au niveau individuel. Cela nous permet de donner une mesure quantitative des écarts entre les deux sources, une tâche qui est difficile compte tenu de la nature différente des objets décrits par les données.

Contents

Remerciements (Acknowledgements) **i**

Abstract **iii**

Table of contents **xiii**

List of Figures **xix**

List of Tables **xxiii**

List of Acronyms **xxv**

Publications **xxvii**

1 Introduction **1**

2 Context and tools **7**
2.1 Data forms . . . 7
2.1.1 Geographical representations and transportation networks . . . 7
2.1.2 Population flows and individual chains . . . 8
2.1.3 Trajectories . . . 12
2.1.4 Link counts . . . 13
2.2 Data sources . . . 15
2.2.1 Surveys . . . 15
2.2.2 Passive data . . . 16
2.2.3 Using mobile phone data . . . 20
2.3 Mobility models . . . 24
2.3.1 4-steps models . . . 25
2.3.2 Activity-based models . . . 26
2.3.3 Data requirements . . . 27
2.4 Respecting privacy . . . 28
2.4.1 Pseudonymization is not enough . . . 29
2.4.2 Anonymization of location chains . . . 30
2.4.3 Anonymization of other mobility data . . . 31
2.5 Conclusion . . . 32

3 Anonymisation of OD matrices **35**
3.1 What is anonymization and why it is difficult . . . 35
3.1.1 The various definitions of anonymous . . . 36
3.1.2 Anonymization as a combinatorial optimization problem . . . 37
3.2 State of the art of anonymization . . . 40
3.2.1 Trajectory anonymization . . . 40
3.2.2 Generalization and suppression for relational data . . . 41
3.2.3 Differential privacy . . . 43
3.2.4 Positioning of our anonymization method . . . 46

3.3 Anonymization methodology . . . 47
3.3.1 Overview . . . 47
3.3.2 Best pruning problem . . . 49
3.3.3 Pruning problem with hard suppression constraint . . . 51
3.3.4 Pruning problem with soft suppression constraint . . . 56
3.3.5 Global suppression constraint . . . 60
3.3.6 Proposed models . . . 60
3.3.7 Parameter choices . . . 61
3.4 Data . . . 61
3.5 Anonymization benchmark . . . 63
3.6 Discussion on the performance indicators . . . 66
3.7 Results . . . 67
3.8 Conclusion on the anonymization . . . 70

4 Synthetic population with activity chains 73
4.1 Introduction . . . 73
4.2 Generating a synthetic population . . . 74
4.2.1 Socio-economic features . . . 74
4.2.2 Integerization . . . 80
4.2.3 Activity chain assignment . . . 81
4.2.4 Evaluating a synthetic population . . . 82
4.2.5 Positioning . . . 83
4.3 Data used for the synthesis . . . 83
4.3.1 Zoning . . . 84
4.3.2 OD matrices from mobile data . . . 84
4.3.3 National census . . . 85
4.3.4 Household Travel Survey (HTS) . . . 86
4.3.5 Commute matrices . . . 88
4.3.6 Synthetic population with activity chains . . . 88
4.4 Temporal calibration . . . 90
4.4.1 Problem statement . . . 91
4.4.2 Formalization as an optimization problem . . . 92
4.5 Results . . . 96
4.5.1 Rescaling coefficients . . . 96
4.5.2 Socioeconomic composition of the population . . . 98
4.5.3 Number of trips by hour: . . . 99
4.5.4 Distribution of agenda popularity: . . . 99
4.6 Discussion and conclusion . . . 101

5 Spatializing synthetic travel demand with OD matrices 103
5.1 Introduction . . . 103
5.2 Spatialization background . . . 104
5.2.1 Primary activity locations . . . 104
5.2.2 Secondary activity location . . . 104
5.2.3 Mobile phone data for synthetic travel demand . . . 106
5.2.4 Evaluating synthetic mobility . . . 107
5.2.5 Positioning . . . 108
5.3 Spatialization methodology . . . 109
5.3.1 Problem statement . . . 109
5.3.2 Spatialization . . . 110
5.4 Experiments and results . . . 114
5.4.1 Assessment metrics . . . 115
5.4.2 Distribution of distances . . . 115
5.4.3 Matching with the OD matrices . . . 115
5.4.4 Matching with the commute matrices . . . 117
5.4.5 Discussion . . . 119

5.5 Conclusion on the spatialization . . . 120

6 Conclusion and research perspectives 123
6.1 Conclusion . . . 123
6.1.1 Anonymization of mobile data . . . 124
6.1.2 Mobile data for synthetic travel demand . . . 124
6.2 Perspectives . . . 125
6.2.1 Better coupling origin-destination in anonymization . . . 126
6.2.2 Hierarchical population . . . 126
6.2.3 Better estimation of factor graphs . . . 127
6.2.4 Better matching emission maps from mobile data . . . 127
6.2.5 Better taking into account commute matrices for primary locations . . . 129
6.2.6 Additional conditioning on the spatialization drawing . . . 130
6.2.7 Better conciliation of mobile data and HTS . . . 130

Bibliography 131

7 Appendices 151
7.1 Optimization and duality . . . 151
7.1.1 Solving a primal problem . . . 151
7.1.2 Relaxation and dual problem . . . 153
7.1.3 Formulations of primal and dual problems . . . 155
7.2 Probabilistic graph models . . . 157
7.2.1 Bayesian networks . . . 157
7.2.2 Undirected graphical models . . . 161
7.2.3 Limitations . . . 162
7.3 Detailing of the anonymization results . . . 162
7.4 Temporal analysis is still possible even though separate OD-matrices are generalised differently for each timestep . . . 163
7.5 Tweaking of the probability law given by the factor graphs . . . 163
7.6 Preprocessing of the agendas . . . 166

List of Figures

2.1 Left: train, subway and road infrastructures around Paris, France. Right: representation of the infrastructure as a multimodal network. Illustration inspired from Asgari et al. (2016) 8
2.2 Example of the zoning used for statistical purposes around Paris, France. 9
2.3 Attraction and emission of yellow taxi flows for an average morning of January 2019 in Manhattan. Data from New York City (TLC, 2019). Zones that attract (respectively, emit) more trips are in red. 10
2.4 Left: OD matrix of an average morning of January 2019 in Manhattan. Right: OD matrix of the 15th January at 10am. Each pixel corresponds to a trip from an origin zone (associated to the row) to a destination (associated to the column). In gray, couples for which no trip has been observed. Note how some zones stand out as very active while others neither attract nor emit trips. Note also that as the temporal precision increases, the OD matrix becomes very sparse. . 11
2.5 Voronoï diagram of the cellular network of Senegal of telecom provider Orange in 2013, as described in the D4D data challenge (de Montjoye et al., 2014). Raster: OpenStreetMap. . . . 21
2.6 Detection of stay points with TRANSIT (Bonnetain et al., 2021). 23
2.7 A data set of trajectories that is $k^{\tau,\epsilon}$-anonymous for trajectory i. Example inspired from Gramaglia et al. (2017). 32

3.1 Example of generalization and suppression of a simple OD matrix. Note that the flows here have been clustered without any kind of constraint. Some approaches seek particular solutions, notably ones where the generalizations form a partitioning of the domain. 38
3.2 Example of a spatial generalization hierarchy. The root represents the whole study map, and the children of a node form a partitioning of the parents. An individual present in area A in the data can be generalized to be shown as present in area B, C or D depending on what is necessary in order to hide them in a group of k individuals. 39
3.3 Merging two generalized trajectories (one blue on the left, one magenta on the right) in Glove. Each trajectory is defined by its generalized points, which appear as boxes containing the actual spatio-temporal points. The resulting generalized trajectory is given by the three transparent boxes. 41
3.4 Left: generalization lattice of a OD-matrix where the maximum level is 3. The first number represents the level of generalization of origins, the second represent the level of generalization of destinations. The original OD-matrix, without generalization, is represented by the node (1 1). Right: Successive generalization steps drawing a path in the lattice, where each direction can be either chosen by an heuristic as in Datafly, or by an exact algorithm as in OIGH. 42
3.5 Example of an OD matrix with hierarchies for origins and destinations, and one possible output of uniform generalization. 42
3.6 Left: initial OD matrix to anonymize. Right: same flows with their generalized areas of origin and destination. 43
3.7 Distribution of the Laplace law of mean 0 and scale 1. 44
3.8 Probability of the noised volume for 4 different values of true volume. Reading: the probability that a flow of volume 1 is shown as a flow of volume 0 after the noise process is around 0.3, while it is only around 0.1 if the true volume is 2. 45
3.9 Classification of k-anonymization approaches with respect to the space of solutions they consider. 47

3.10 Left: Example of a complete generalization hierarchy. Middle: Example of a pruning of the hierarchy, satisfying the tree constraint. Right: Interpretation of the pruning as a partitioning of the values. . . . 48
3.11 Left: example of an OD matrix with the hierarchy over the origins. Middle: emission flows from each initial and generalized origin. Right: corresponding penalty obtained by setting $\pi_n = \left(v_{target} - \sum_{d \in leaves(T)} v_{n \to d}\right)^2$ with $v_{target} = 10$. . . . 51
3.12 Left: example of an OD matrix with the hierarchy over the destinations. Middle: aggregation penalty for each destination for one given origin. Right: suppressed volumes for each destination, considering $k = 10$. . . . 52
3.13 Left: General shape of the lagrangian dual function of problem D. Right: Geometrical interpretation of one step of the algorithm SBA used to find the maximum. . . . 56
3.14 Left: separately generalizing the destination map of each origin. Right: solving for the global problem under the unified constraint C . . . 60
3.15 All generalization levels considered by OIGH for a hierarchy that does not have all leaves at the same depth. . . . 65
3.16 Comparison of the performance of our ATG-Dual approach versus ATG-Soft, Glove-sk, Glove, OIGH, and Mondrian. Each box represents the distribution of the performance over one data set, expressed as the ratio of the benchmark over ATG-Dual. First column: comparing generalisation error G. Second column: comparing the reconstruction loss E. Third column: comparing the distribution distance L_1. Fourth column: comparing the computing times. In each case, the line $y = 1$ represents matrices for which the benchmark has the same performance as ATG-Dual, and the lines $y = p$ represents matrices for which the benchmark is p times worse than ATG-Dual. The box plots correspond to the data sets in this order: `nyc`, `civ`, `senegal_crop`, `senegal`, `senegal_split`, `senegal_big`. . . . 68

4.1 IPF setting for a population with 3 features. The population is represented by a 3-dimensional tensor where each entry is the number of people for a given joint value. The target totals are given by vectors. For any given feature, its target total vector can be matched by applying the appropriate scaling factors to each slice (for example the yellow slice for the second value of feature 1). As we proceed to calibrate the population with respect to feature 2, the scaling factors associated to the pink cells can be modified, but the algorithm eventually converges. . . 75
4.2 Illustration of the quadratic function, the raking function, and the logit function used for the calibration problem. The bounds are $(l, u) = (0.1, 5)$. . . . 78
4.3 Hierarchical model proposed by Sun et al. (2018) The plate notation indicates that a random number N of persons are independently drawn once the household type is decided. . . . 80
4.4 Bayesian network proposed by Joubert and De Waal (2020) to model both the population and the agendas, learned by tabu search. . . . 82
4.5 Positioning of the population synthesis pipeline and the steps discussed in this chapter. . . . 84
4.6 Map of the partitioning of the study area of the city of Lyon. Background: OpenStreetMap . . 85
4.7 Example of merging a short activity (Shopping) with the next one. The least important activity is discarded. If the inbound trip is the longest of the two, then it is considered as the trip going from work to home (b). Else, the outbound trip is considered (c). . . . 87
4.8 Emission and reception map of the commute matrix for Lyon, France. . . . 89
4.9 Commute matrix of Lyon. . . . 90
4.10 Comparison of the trips performed by the synthetic population and the mobile phone data, in actual volumes and in distribution. . . . 91
4.11 Histogram of the rescaling coefficients obtained by our various approaches . . . 97
4.12 Lorenz curve of the rescaling coefficients obtained by our competing approaches. If we note AUC the area under the curve, then the Gini coefficient is given by $1 - 2 \times AUC$. . . . 98
4.13 Matching of the socioeconomic marginals of the synthetic population (in red) compared to the HTS (in green). . . . 100
4.14 Distribution of trips taken during the day. For each hour, the first bar in orange is the fraction observed in the mobile data, the second in magenta is according to our synthetic demand, the third in blue is according to the reference travel demand, and the last in yellow is according to the HTS. . . . 101

4.15 Popularity of agendas in our synthetic demand vs. popularity in HTS. The top-right dot is the empty agenda. . . . 101
4.16 Cumulative distributions of the agendas from most to least popular, in our synthetic demand, the reference travel demand, and HTS. . . . 102

5.1 Illustration of a space-time prism between two anchor points $A = (x_A, y_A, t_A)$ and $B = (x_B, y_B, t_B)$, describing all the possible spatio-temporal points to start an activity of duration d. $B' = (x_B, y_B, t_B - d)$ is the most late point at which it is possible to start the activity without being late to meet anchor point B. Figure inspired from Lenntorp (1976). . . . 104
5.2 Representation of the force model for the attribution of secondary locations. The grey locations are fixed. The direction of the arrows indicates the direction of the spring force, while the thickness indicates its strength. . . . 105
5.3 Positioning of the spatialization approach presented in this chapter. . . . 108
5.4 Naive assumption of the dependency of the locations on an example chain "h-w-h-w-o-h". . . . 110
5.5 Dependency assumption of the linked forward approach. . . . 111
5.6 Locations distribution with explicit links between states sharing the same location for activity chain "h-w-h-w-o-h". . . . 112
5.7 Factor graph representing the general factorization of Fig. 5.6. . . . 112
5.8 Sub-factorization of a factor as three factors of two variables. . . . 112
5.9 factorization of the location distribution with all assumptions. White squares denote factors that are conditional probabilities. Black squares are equality factors, forcing their neighboring states to be equal. . . . 113
5.10 A simple commute agenda "h-w-h". . . . 114
5.11 Distribution of trip lengths (in meters) between zones 116
5.12 Scatter plot of the OD matrices from synthetic demands compared to the mobile data. . . . 118
5.13 Scatter plot of the OD matrices from synthetic demands compared to the official commute matrix. 119

6.1 Distribution of the distance of commute trips and non-commute trips in our HTS. . . . 129

7.1 Examples of an unbounded and a bounded linear programs. Constraints are represented as lines where it is allowed to be on one side only. The intersection of the constraints is a polygon in which the objective function is represented as a color gradient. . . . 152
7.2 Example of Bayesian network. . . . 157
7.3 Markov blanket of node X_4 . . . 158
7.4 A possible junction tree for the Bayesian network illustrated in Fig. 7.2. Example taken from Nuel (2012). . . . 159
7.5 An undirected graph that can actually represent several possible factorizations. . . . 162
7.6 Two factorizations that are represented by the same undirected graph but by distinct factor graphs. . . . 162
7.7 Weekly profiles of $o \to d$ flows of various importance in dataset nyc. Blue curve: profile obtained from the OD-matrices anonymised by ATG-dual then reconstructed. Orange curve: profile obtained from the initial OD-matrices. . . . 164
7.8 Factorisation of the location distribution with all assumptions. White squares denote factors that are conditional probabilities. Black squares are equality factors, forcing their neighboring states to be equal. Note that there are two factors between Z_1 and Z_2 as a result of the factorisation described in Section 5.3.2 . . . 165
7.9 Logic of each step of the pre-processing of the HTS agendas. Going through each trip in order, we decide if it must be discarded, added to the agenda as is, merged with the previous trips of the chain, or kept as a **candidate** to be merged with the next trips of the chain. . . . 168

List of Tables

2.1 Examples of individuals defined by their socio-economic variables. 12
2.2 Examples of activity chains. Each row is an activity where we specify the start time of the trip to go there, the transport mode, and the purpose. The first activity of the day, activity n°0, does not have an associated trip and is generally "home". 13
2.3 Example of a k^m-anonymous data set with four variables, with $k = m = 2$. 31

3.1 Flows of Fig. 3.6 and their non-homogeneous generalization. Note how any particular initial flow has its origin and destination included in enough generalized flows so that the cumulated volume is above $k = 10$. 43
3.2 Descriptive statistics of the data sets used for the experiments (#matrices: number of matrices available in the whole data set, where each matrix represents the flows over a time step; #tiles: number of initial tiles over which the matrices are set; density: average graph density of the matrices; avg. #flows: average number of flows among the matrices in the data set; avg. vol.: average sum of the flows; %anon. flows: fraction of flows that are above $k = 10$; %anon. vol: fraction of individuals that are in flows above $k = 10$). Note that due to performance concerns, we do not evaluate the benchmark on all available matrices and we rather choose a small sample of matrices for each data set. 63
3.3 Best v_target for the studied data sets for various values of k. 63
3.4 Performance on samples of the data set senegal_crop, nyc, civ, and senegal. $\bar{G}$: mean generalisation error (Eq. 3.6); E: normalized reconstruction loss (Eq. 3.5); S: fraction of volumes suppressed (Eq. S); L_1: distribution distance (Eq. 3.7). The reported time is the total computing time (in seconds) to run the anonymisation of all matrices of the samples. Note that differential privacy adds volumes, as it mostly applies a positive noise on a sparse matrix. 70
3.5 Performance on samples of the data sets senegal_big and senegal_split. 70

4.1 Example of IPU to rescale a population of households containing persons. Example from Ye et al. (2009). 76
4.2 Example of a synthetic agent requiring an activity chain. 82
4.3 Example of HTS respondents. The matching attributes are highlighted. The last row indicates the number of match that the agent in Tab. 4.2 have when considering the n left-most attributes. For example, if we consider only `age` and `gender`, then there are 3 matches. 83
4.4 Density of the transition matrices with respect to the time step. 86
4.5 National census variables used for the synthetic population. 86
4.6 HTS trip variables describing the activity chains 86
4.7 Impact of the preprocessing on the HTS 88
4.8 Variables used in the statistical matching to assign activity chains from HTS to agents from census 89
4.9 Notations used in the formalization of the temporal calibration problem. 95
4.10 Evaluation of the various rescaling approaches for the temporal calibration. 99

5.1 Classification of the spatialization models used in this study. 114
5.2 Fit between the observed OD matrices and the synthetic travel demands of our benchmark. . . 117
5.3 Fit between the official commute matrix and the synthetic travel demands of our benchmark. . 120

7.1 Detailed results 169

List of Acronyms

ATG	Adaptative Tree Generalization
CDR	Call Detail Record
GDPR	General Data Protection Regulation
GPS	Global Positioning System
HTS	Household Travel Survey
INSEE	Institut National de la Statistique et des Études Économique
IPF	Iterative Proportional Fitting
IPU	Iterative Proportional Updating
MCMC	Markov Chain Monte Carlo
MSE	Mean Square Error
NSD	Network Signalization Data
OD	Origin-Destination
PPDP	Privacy Preserving Data Publishing
SBA	Some Breakpoints Algorithm
VGH	Value Generalization Hierarchy

Publications

This thesis has been the opportunity for multiple publications, which are listed below.

International peer-reviewed journals

Benoit Matet, Angelo Furno, Marco Fiore, Etienne Côme, and Latifa Oukhellou. Adaptative generalisation over a value hierarchy for the k-anonymisation of origin–destination matrices. *Transportation Research Part C: Emerging Technologies*, 154:104236, 2023a. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2023.104236. URL https://www.sciencedirect.com/science/article/pii/S0968090X23002255

International conferences

Benoît Matet, Etienne Côme, Angelo Furno, Loïc Bonnetain, Latifa Oukhellou, and Nour-Eddin El Faouzi. A lightweight approach for origin-destination matrix anonymization. In *29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 487–492, 01 2021a. doi: 10.14428/esann/2021.ES2021-56

Benoît Matet, Etienne Côme, Angelo Furno, Latifa Oukhellou, and Nour-Eddin El Faouzi. Mobile phone origine-destination matrix anonymization and analysis. In *Proceedings of the Conference of Complex Systems*, page 130, 10 2021b. URL http://sci-web.net/CCS2021/CCS2021_BookOfAbstracts.pdf

Benoît Matet, Etienne Côme, Angelo Furno, Sebastian Hörl, and Latifa Oukhellou. Use of origin-destination data in synthetic travel demand synthesis. In *10th International Symposium on Transportation Data & Modelling (ISTDM2023)*, pages 23–26, 06 2023b. doi: 10.2760/135735

Benoît Matet, Etienne Côme, Angelo Furno, Sebastian Hörl, and Latifa Oukhellou. Use of origin-destination data for calibration and spatialization of synthetic travel demand. In *11th symposium of the European Association for Research in Transportation*, 09 2023c. URL https://transp-or.epfl.ch/heart/2023/abstracts/hEART_2023_paper_4604.pdf

Under submission

Benoît Matet, Etienne Côme, Angelo Furno, Sebastian Hörl, Latifa Oukhellou, and Nour-Eddin El Faouzi. Improving travel demand synthesis using origin-destination matrices from mobile phone data. under submission, 11 2024

Chapter 1

Introduction

Starting from the beginning of this century, more than half of the world's population lives in urban areas (United Nations, 2018). This number is expected to increase to two-thirds by 2050, with about one-third of urban dwellers living in cities with more than 1 million inhabitants. The proper organization of mobility within these cities is a key determinant of air and noise pollution, accessibility, and the overall quality of life of their inhabitants. The social costs of inappropriate urban transport networks are also associated with economic costs. In Europe, it is currently estimated that urban congestion costs a total of 110 billion per year (Brannigan et al., 2017).

A necessary information for the design of an adapted transport network is the knowledge of people's needs. The study of people's mobility behavior, regardless of the specific aspect that is measured, is grouped under the umbrella term **travel demand**. Some purely pragmatic, for-profit applications of travel demand studies would be to identify the best places to display advertisements, or to decide the new locations for a chain store. Other, more public-interest applications aim to monitor and understand the impact of public policies. A more thorough understanding of the dynamics of the system can also help identify how to create desirable behaviors of the transportation networks: we might want to encourage the system in various directions, such as minimizing the total time spent in traffic jams, minimizing the carbon footprint of trips, maximizing the use of bicycles, encouraging car users to switch to public transportation, or any other criterion we think is beneficial to the area or its population.

This thesis is motivated by the need to design new and sustainable transportation networks in modern cities, that replace the preponderance of cars with a resilient and much more efficient public transportation network. Transport infrastructures, especially railways, are remarkably expensive: the average French tram costs 34 million euros per kilometer (Bensalah et al., 2017), while the excavation costs drives the prices of metro lines even higher. Expressed in 2015 US dollars for comparison, the 1999 Jubilee Line Extension in London cost $505 million per kilometer, while the 2015 Flushing Line Extension in New York, a city known for its prohibitive subway construction costs, was estimated at $1.07 billion per kilometer (Hess, 2016). The 205 kilometers of the project Grand Paris Express, the largest urban transit project under construction in the world, is expected to ultimately cost around 42 billion euros, for an average in US dollars of around $220 million per kilometer (Commission des finances,

2020).

It appears then as a very basic requirement to be able to understand in detail the mobility behavior of the population concerned, so that the resources put in infrastructure are used in the most efficient way. However, to this day such knowledge is limited to aggregate, static indicators, and no expert or official can, for example, give the exact number of people who went from one point in the city to another, yesterday at 10am. In fact, the understanding of the urban mobility is so limited that some counter-intuitive results are yet to be explained. In a growing number of documented cases, it has been observed that closing a road does not generate more traffic in the neighboring area (Cairns et al., 2002). What happens to the former road users is still unclear, to the point where the phenomenon is referred to as "traffic evaporation" (Nello-Deakin, 2022).

There are two reasons why the exact whereabouts of a population are difficult to estimate and understand. First, because cities are complex systems in which people, organizations, and vehicles interact in diverse and intricate ways, modeling a transportation system requires modeling the decisions of its users, their preferences, and the influences they have on each other. In the words of physicist Murray Gell-Mann, "Think how hard physics would be if particles could think". Second, detailing people's whereabouts is generally frowned upon in a world where privacy is considered a fundamental right. Had he been more sensitive to these concerns, Murray Gell-Mann might have added: "And how hard it would be if particles didn't want you to measure them".

Various data sources are available to measure parts of this mobility behavior. They are all incomplete in their own way and generally measure different objects in different contexts, making their joint use difficult. Merging these data sources to understand the details of urban traffic, while ensuring the privacy of the available sources, has been the guiding principle of the work presented here.

The most obvious way to measure the mobility of a large portion of the population would be to access the GPS of mobiles phones. This first possibility exists, and gives what is called Location Based Services (LBS) data. LBS data is issued from mobile applications (the services) that probe the location of the device, generally for operational reason, and transmit it to the data collector for additional processing. It can yield interesting results, but is often limited in its reach, especially if it only tracks users of a specific application. The second most obvious source would be the ticketing of public transportation: in most modern subway systems, it is required to validate a personal pass upon entering and leaving the subway, making it possible to collect what are called **Origin-Destination (OD) matrices** of travelers. However, ticketing data can only monitor the public transport usage, not to mention that it is often not required to validate upon leaving buses and trams, leaving us with only a map of where the traveler get on.

Anterior to these modern sources, travel demand studies relied simply on surveys, asking the respondents for some personal information and most of all a detailed report of one day of their week. Such transport-oriented surveys are called **Household Transportation Surveys (HTS)** throughout this work. They are invaluable both by their level of detail and their handling of individual coherence: rather than simple flows described by OD matrices, HTSs describe the complete agenda of individuals for a given day, with the location, purpose, and

hours of each activity, but also the transport mode used between the activities. This allows to model the interactions between the activities and between the individuals, which is the cornerstone of modern mobility models. In fact, mobility models are designed to take as input data that resemble an HTS but describing a complete population. Such data needs to be synthesized, hence its name of **synthetic travel demand**. It is very hard to create comprehensive synthetic travel demand solely from an HTS, as they are limited by their very low number of respondent, aggravated by their high no-response rate. This makes them unreliable for the estimation of spatial locations, as only a handful of the locations in the city will actually be visited by one of the HTS respondent. From a statistical perspective, if the city is coarsely divided into 200 zones, the number of possible combinations of locations for a given chain of 5 activities is 320 billions, and there is no way a survey can give an reliable estimate of the likelihood of each of these possibilities.

One possible source that has recently come under attention to measure the movements of the population is **mobile phone data**. Contrary to GPS traces captured by mobile application, what we call mobile phone data in this work is the operational data collected by mobile operators. This includes two sources: **Call Detail Records (CDRs)**, which are collected for billing purposes, and **Network Signalization Data (NSD)**, which is all the messages exchanged during the normal operation of the network. Both sources log, among others, a timestamp, the identifier of a cellphone antenna, and the identifier of the SIM card of a mobile user, indicating that the user was likely in the vincinity of the antenna at the indicated time. The main advantage of mobile phone data is its extended reach: almost everybody has a phone, and the market is generally divided into few distinct operators, meaning that an operator can easily gather information on a significant fraction of the population. As such, mobile data constitutes a promising asset toward an exhaustive understanding of urban mobility. Their use raises hopes for the ability to one day understand travel demand with cheap and fast approaches, capable of producing insights into the dynamics of transportation with little delay, and eventually to anticipate travel demand in hypothetical situations. However, before these hopes can be fulfilled, two main issues need to be addressed:

- The first issue is the problem of precision and correctness. The precision of mobile data is limited to the indication of a cellphone antenna, whose actual coverage area is generally unknown, impractical to use, and large compared to the distance separating different points of interest in a city. The temporal precision of CDRs is widely admitted to be unsatisfactory due to the low sampling rate, a problem that is mitigated but not entirely resolved by NSD (Bonnetain, 2022). In addition, there is no detail other than spatio-temporal points, and we could say that the precision in terms of transport mode or socio-economic profiles is null. Mobile data can also be incorrect, in the sense that the information, even if precise, is factually false. The actual logic governing the connection of mobile phones to network antennas depends on multiple, uncontrolled factors, and the technical specifications for phone-antenna communications are aimed at maximizing the quality of service, but not the ease of use for mobility studies. This results in noisy, chaotic trajectories with missing trips or conversely artifact movements. Directly interpreting the behavior observed from the mobile data as representative of the population also leads to inevitable errors, as each mobile operator's customer base is skewed relative to the population as a whole.

- The second issue is of course the problem of privacy, because despite the problems mentioned above, mobile data is first and foremost a promising tool to monitor the movements of a population on an unprecedented scale without any intervention by the targeted individuals. It is only natural to expect that such a rich data source should be completely anonymous, to the point where it could be publicly released. After all, they are records of the public life, put to use to help develop public policy. This privacy liability is also an obstacle to the dissemination of the data. In european countries, the particularly restrictive GDPR (European Commission, 2018) is a strong disincentive for mobility data owners to share any form of data, even simple OD matrices. This hampers the research efforts aimed at developing better methods of transportation planning for greener, safer, and more resilient transportation networks. Note that this problem is not inevitable: The GDPR aims to protect personal information, while transportation studies are normally more interested in general mobility trends.

The main aim of this work is to promote safe mobility studies using mobile data. To this end, we study particularly the use of OD matrices, as they are by nature more prone to being fully anonymized than complete mobile trajectories. We consider that OD matrices derived from mobile phone data are a good opportunity to introduce a data-driven approach to the spatialization of synthetic travel demand. Mobile data provides extensive, dynamic, spatio-temporal information on the whereabouts of the population without individual details, while HTSs provide rich individual descriptions that lack spatial representativeness. Combining both sources is an opportunity to synthesize a comprehensive population that is realistic at the individual level while also behaving realistically at the population level. These considerations lead to two main research questions:

- **Question 1:** How can we guarantee the privacy in OD matrices while preserving their valuable information?
- **Question 2:** How can we use OD matrices derived from mobile data to generate more realistic and comprehensive synthetic travel demand?

Before answering these questions, we begin in Chapter 2 by an overview of the context and the various data used in transportation, along with introductions to the mathematical tools that will come useful in this work. Then, we address the question of anonymization in Chapter 3. The significant size (*i.e.,* number of distinct flows) and high number of modalities (produced by a high-resolution zoning) of OD matrices call for an adapted, fast algorithm that can efficiently anonymize them. In this work, we develop a lightweight approach for the k-anonymization of OD matrices that exploits the low dimension of the data to explore a larger solution space than regular generalization algorithms, while keeping relevant restrictions of the search space in order to be scalable on matrices with high number of flows. We apply it to a variety of real-world large-scale OD matrices collected by the New York City Taxi and Limousine Commission and derived from the Data for Development (D4D) challenge organised by Orange in Senegal and Côte d'Ivoire. Compared to an extensive benchmark composed of regular generalization algorithms as well as mobility anonymization approaches, we show that our method is 27% more precise and 9 times faster than comparable approaches able to scale on the same data sets.

In Chapter 4, we provide an overview of the problem of travel demand synthesis with and

without the use of mobile data. We focus on the state-of-the-art pipeline introduced by Hörl and Balac (2021), which is based on a succession of steps to integrate various data sources. As a first contribution in this regard, we propose a method for the temporal calibration so that the temporal distribution of trips made during the day match the distribution observed from the mobile data.

Then, in Chapter 5, we study the use of OD matrices in the problem of spatialization of the travel demand. We develop a spatialization step replacing *ad-hoc* approaches, that models location chains as a probabilistic graph model whose parameters are estimated from the OD matrices. Compared to other graph-based approaches, our work integrates the spatialization into the more general synthesis pipeline described by Hörl and Balac (2021). We also discuss a methodology to estimate these models, as their general, non-tree structures prevent us from directly interpreting the OD matrices as transition probabilities. We propose a variety of graph structures which offer different trade-offs between recreating the OD matrices at the population level and respecting realistic agendas at the individual level. This allows us to give a quantitative measurement of the discrepancies between the two sources, which are otherwise hard to measure given the different nature of the objects described by the data.

Finally, in Chapter 6 we conclude this work and summarize the perspectives of future research toward comprehensive, privacy-preserving transportation data.

Chapter 2

Context and tools

In this chapter, we give an overview of the context of travel demand, the data, the models, and the tools that we will use. We describe the different forms of data that are commonly used in transportation research in Sec. 2.1, and the sources that provide such data in Sec. 2.2. We then present how they are used in travel demand models in Sec. 2.3, and we discuss the privacy risks that arise with the processing of such data in Sec. 2.4. This chapter is completed by Appendix 7.1 and 7.2, which present the formal tools that are used throughout this work with a focus on the logic and the approaches that are privileged when using them. In Appendix. 7.1, we give a brief overview of the field of optimization and the use of duality in the solving of problems, and in Appendix. 7.2 we introduce the theory of probabilistic graph models and the various problems associated with them.

2.1 Data forms

In this section, we review the various forms in which transportation data can come, each different form corresponding to a different object that is described.

2.1.1 Geographical representations and transportation networks

What we call a transportation network in natural language can actually be represented as a mathematical network (Aleta et al., 2017; Alessandretti et al., 2023). To analyze mobility behavior at the most fine-grained level, we can rely on a network where each edge is a road segment and each node is an intersection between two roads. In such a network, the subway would be represented by another parallel network, whose nodes are stations and are connected to nodes of the surface network. Even though the subway stations are geographically located at the same place as the corresponding node on the surface network, it is better to represent them as separate. This allows the edges to store information about how long (or how expensive) it is to travel from the surface network to the subway. Similarly, even though other transport modes technically share the same road segment, it is better to represent each mode by a separate graph with only sparse connection to the others, which contain the information about the cost of changing modes in the same way that the edges of a single network convey information about the cost of travel (Asgari et al., 2016). An illustration of such a network is given in Fig. 2.1.

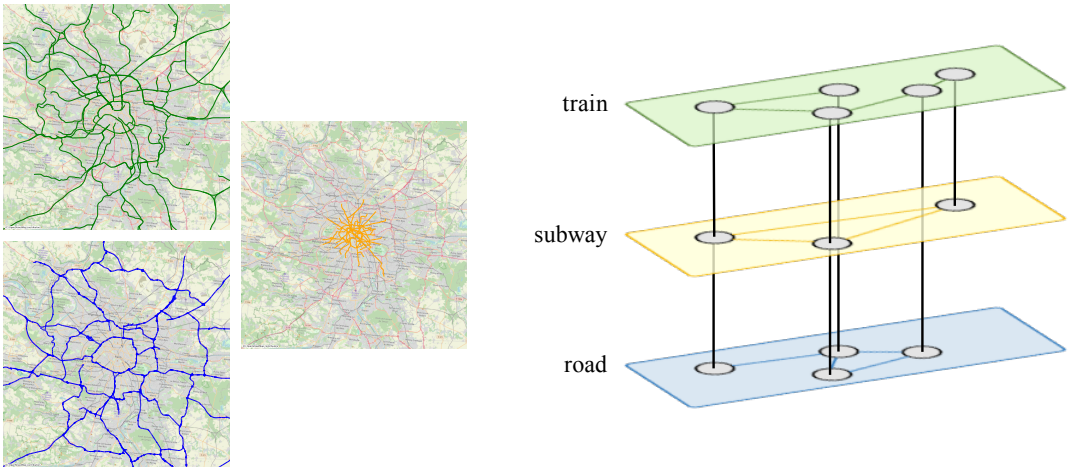


Figure 2.1: Left: train, subway and road infrastructures around Paris, France. Right: representation of the infrastructure as a multimodal network. Illustration inspired from Asgari et al. (2016)

Other scales can be considered for transportation analysis. A less detailed description of the city can be a division into zones, that can be either square tiles or administrative zones (Hagen-Zanker and Jin, 2015). An example of the IRIS statistical areas used by French statistical institute INSEE for the city of Paris, France, is given in Fig. 2.2.

In such a map, the transport infrastructure can still be included in the form of edges, indicating the cost of going from one zone to another depending on the transport mode. Within the zones, Points of Interest (POIs) can be identified, *e.g.,* shops, schools, hospitals or offices, which can be interpreted as the actual precise locations that individuals go to when they are in the zone (Liu and Long, 2016). Some works of research make use of more detailed descriptions of the transportation network. When analyzing the road network of a neighborhood, one can consider the different lanes of the roads on which cars can overtake each other, or the distribution of red light queues. These works are not so much interested in modeling travel demand as they are in studying road and intersection designs that maximize vehicle flows in rather restricted study areas (Lebacque, 2005; Bani Younes and Boukerche, 2016; Natafgi et al., 2018). In even more detailed cases, roads are not represented as edges but as spaces on which people and cars can move in any direction (Gorrini et al., 2018; Kathuria and Vedagiri, 2020). These studies are not interested in travel demand, but rather in the interactions between vehicles and pedestrians to assess the safety of a given intersection.

2.1.2 Population flows and individual chains

The location of transport users in a region can be studied with different levels of detail: in some cases the administration of a territory is satisfied with a map of the zones of attraction and emission of trips (Fig. 2.3). In other cases, the details of the organization of the flows between these zones are required (Fig. 2.4). Such a detailing of flows is referred to as a **Origin-Destination (OD) Matrix**. This name comes from visualizing the data as a table where each row is an origin and each column is a destination. This particular representation is rarely

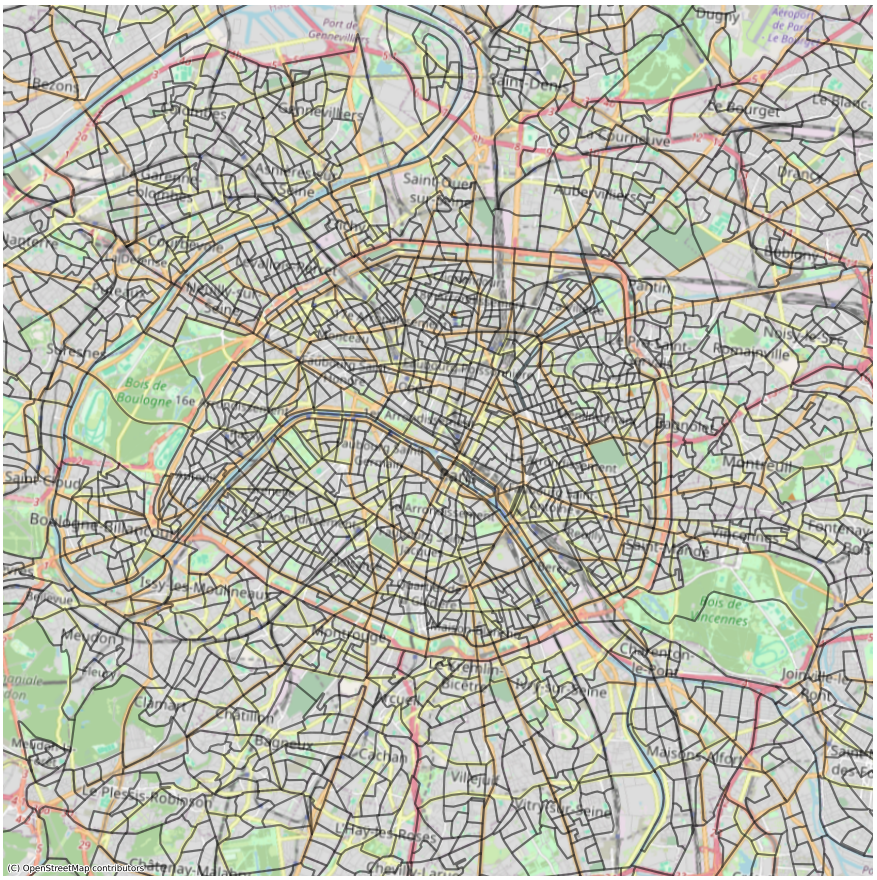
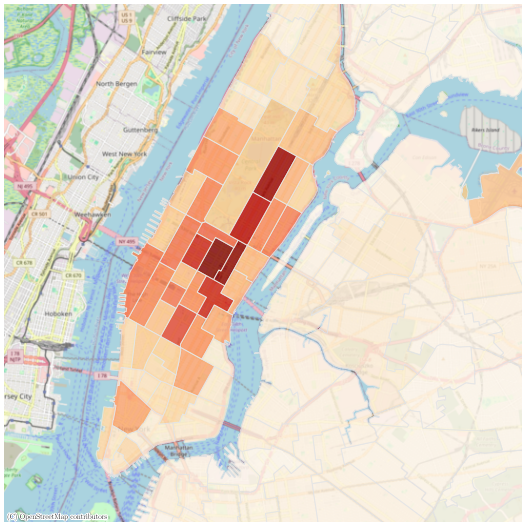


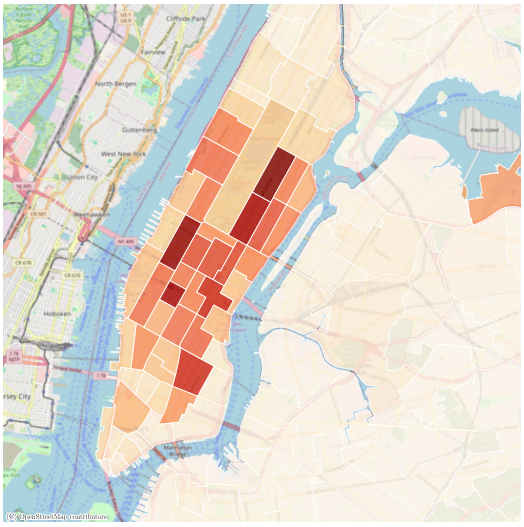
Figure 2.2: Example of the zoning used for statistical purposes around Paris, France.

used by itself because OD matrices typically describe numerous zones, many of which are not actually connected, resulting in a very sparse matrix with a density often below 5%. OD matrices are a good summary to describe the movements of a population and are fairly natural to interpret. Because of this, they are the form of data considered in numerous transportation applications, such as short-term traffic forecasting (Chen et al., 2011; Pamuła and Żochowska, 2023), long-term planning (Wang et al., 2015), or as validation data for other problems such as transit assignment models (Tavassoli et al., 2018).

However, OD matrices remain limited as it is impossible to observe the structure of a particular individual's trips. This information can be valuable, as we can expect that the various trips made by a user of the transportation network are strongly influenced by their other trips, their home location, and other motivations associated with specific spatial or temporal constraints. Thus, the richest form of mobility data that can be available is the details of each trip associated with an identifier to maintain individual consistency. The destination of each trip is expected to correspond to an activity performed by the individual and must necessarily be the origin of the next trip. For this reason, in this work we refer to this type of detailing of mobility data as **location chains**. What we will call location chains must describe at least the list of visited locations for each individual in the study. When they describe more information on the actual activities that are performed, they are usually called **activity chains** in the state of the art (Hörl and Balac, 2021). Like OD matrices and emission-reception maps, location chains can describe the spatial information with varying



(a) Map of trip attraction.



(b) Map of trip emission.

Figure 2.3: Attraction and emission of yellow taxi flows for an average morning of January 2019 in Manhattan. Data from New York City (TLC, 2019). Zones that attract (respectively, emit) more trips are in red.

degrees of precision and can also include additional variables to further refine the analysis:

- The **spatial precision**. Studies of a whole country, such as the French traffic forecasting service Bison futé, can be satisfied with information at the level of the region only, but studies within a region must provide more precise information, such as specific municipalities (called communes in France). Studies of a city can consider tiles of a few hundred meters as their zones of interest, or other administrative zoning at the sub-commune level. Note that an OD matrix defined over very fine zones can be expected to be very sparse.

- The **temporal precision**. General trends of attraction and emission zones can be identified with data describing a whole day, but the movements of populations are known to vary widely throughout the day. Rush hours will challenge the overall capacity of the network, while quiet periods can be costly to the community if public transportation is empty. Analyzing these dynamics of the transportation network can require data as accurate as one hour, with some studies going as low as 15 minutes (Barceló and Montero, 2015). OD matrices that contain a temporal information are said to be time-dependent or dynamic, and typically consist of several OD matrices, each associated with a time step and reporting the trips that begin during that time step. Location chains usually include an arrival time at the location and a departure time with a precision to the minute.

- The **transport mode**. Cities have very different possible modes of transportation, each with its own specific infrastructure costly both in terms of money and in terms of living space. Modern trends in European cities tend to discourage cars and encourage soft mobility such as walking or cycling, while we know that different modes imply different destinations and transportation habits (Ribeiro et al., 2020; Thorhauge et al., 2020).

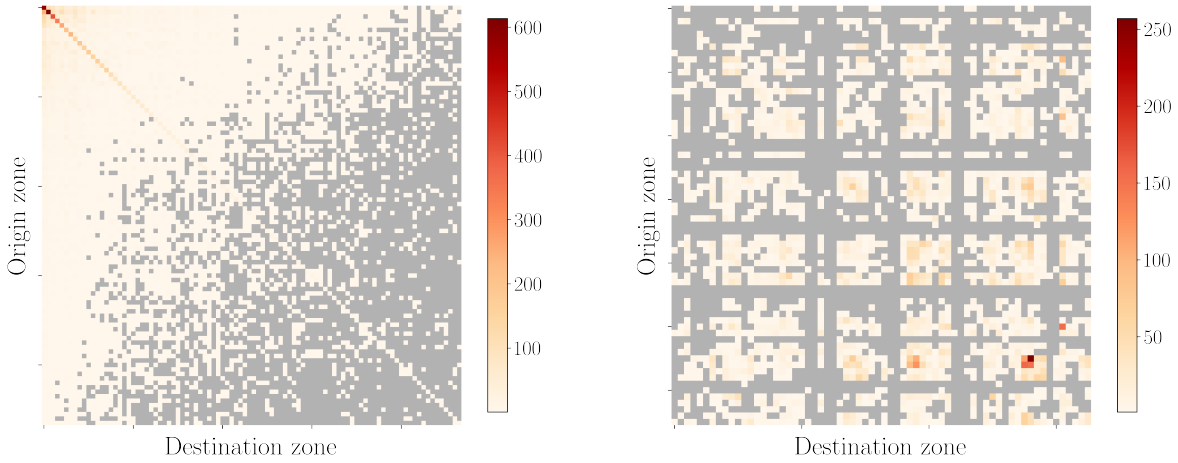


Figure 2.4: Left: OD matrix of an average morning of January 2019 in Manhattan. Right: OD matrix of the 15th January at 10am. Each pixel corresponds to a trip from an origin zone (associated to the row) to a destination (associated to the column). In gray, couples for which no trip has been observed. Note how some zones stand out as very active while others neither attract nor emit trips. Note also that as the temporal precision increases, the OD matrix becomes very sparse.

For these reasons, the ability to specify the transport mode in mobility data becomes a prominent problem. The distinction can be as simple as separating road and rail modes, but urban studies may also be interested in identifying walking and biking trips.

- The **purpose**. Commute trips follow different trends than personal errands, and shopping trips may require a car even though the distance is short. Knowing the purpose of trips can help us better understand how the population appropriates the urban organization and how the different activity zones of a city interact. It also affects the possible interactions we can model between individuals in the same household, since only one person in the household may need to go shopping or take the children to school. Purposes are usually divided into **primary purposes** and **secondary purposes**. The former are generally labeled as "home", "work", or "study" and correspond to activities that must be performed at a fixed location during the day, possibly with time constraints. The latter are activities such as "shopping", "leisure", or other errands, and we assume that the exact time and location of these secondary activities are determined by the individuals to arrange their schedules. Location chains are often detailed by purpose, in which case they are often called **activity chains**. In some works, this term is even used for chains of purposes without location (Joubert and De Waal, 2020).

- The **socio-economic description**. People with different socio-economic profiles have different mobility behaviors and may have preferences for certain transport modes. More importantly, it is necessary to ensure that a transportation network is adapted to the needs of all and does not exclude any particular category of the population. As a result, it is valuable to have a socio-economic description of the people who perform the trips featured in our data. OD matrices almost never include socio-economic description, mostly because they are usually estimated from large scale digital data sources that cannot measure this kind of detail. In contrast, location chains can be derived from

more detailed data sources such as surveys, which are more focused on the individual and can provide a socio-economic description. Common variables for this description are age, gender, occupation, or household income, which are known to impact the mobility behavior (McGuckin and Murakami, 1999; Cheng et al., 2016).

In this work, we mostly use dynamic OD matrices without mode or socio-economic description, and location chains with full description of the trips and the individuals. To compare location chains with OD matrices, it is common to consider the OD matrices implied by the chains. This is done by simply removing the identifier connecting all the trips of an individual. As individual chains tend to feature a very high spatio-temporal precision, it is also advisable to aggregate the resulting OD matrix to a restricted set of zones and time steps in order to obtain flows of more than one person. Sets of locations chains can also be transformed into OD matrices in order to compare them with other location chains data sets (Egu and Bonnel, 2020). Indeed, in its original form, a set of location chains is a realization of a probability law of very high dimension where each realization has a different value, which makes it impossible to perform a meaningful statistical comparison.

An example of the socio-economic description is shown in Table 2.1, with corresponding activity chains illustrated in Table 2.2. Note that the activity chains can contain information about the trips between the activities, such as the transport mode. In that case, in this work we will associate each trip to its destination activity. We will also refer to activity chains as **agendas**, indiscriminately.

Age	Gender	Occupation	Car	Home status	Home zone	Agenda
60+	male	retired	yes	owner	69123	1
30-59	female	executive	yes	owner	6905	2
30-59	male	independant	yes	tenant	6906	3
18-29	female	student	no	tenant	0113	4

Table 2.1: Examples of individuals defined by their socio-economic variables.

2.1.3 Trajectories

In some cases it may be possible to obtain a series of spatio-temporal points that are not to be interpreted as a sequence of activities but only as a sampling of locations, some of which are during a trip and some others during an activity. These trajectories can be obtained from GPS-tracking mobile applications (Hawelka et al., 2014), mobile network usage (Wang et al., 2018), or Bluetooth or wifi sensors (Friesen and McLeod, 2015; Naik et al., 2018). They do not contain any information other than the positions at given times. However, it is possible to infer a semantic interpretation of the trajectories to obtain chains of locations. For this purpose, the general assumption is that when the individuals are static, then it means that they are performing an activity (Bonnetain et al., 2021). The processing of trajectories then requires to detect these **stay points** in order to identify personal points of interest. Identifying stay points can be difficult in itself, as either noise in the trajectories or excessive precision in the data can indicate a movement that is either false or irrelevant in the definition of the activities (Liao, 2020). In the absence of additional information, a time threshold has to be determined above which the stay point is considered as an activity and below which it is

Agenda	Activity	Trip time	Transport mode	Purpose	Location
1	0	—	—	Home	Zone A
1	1	8:00	Car	Work	Zone B
1	2	16:00	Car	Home	Zone A
2	0	—	—	Home	Zone C
2	1	7:00	Bus	Study	Zone D
2	2	18:00	Bus	Home	Zone C
3	0	—	—	Home	Zone E
3	1	9:00	Bike	Shopping	Zone G
3	2	12:00	Bike	Shopping	Zone H
3	3	15:00	Bike	Home	Zone E

Table 2.2: Examples of activity chains. Each row is an activity where we specify the start time of the trip to go there, the transport mode, and the purpose. The first activity of the day, activity n°0, does not have an associated trip and is generally "home".

filtered out as a random hazard of the trip.

The missing information on the trips and activities defined by the stay points can then be inferred (Chrétien et al., 2018). For example, the purpose of an activity can be inferred from the repetition of the stay point and its time of day, which can be indicative of a home or a work activity (Yang et al., 2021). The detection of secondary activities can be much more difficult depending on the precision of the spatio-temporal points, since there is a possible overlap between the duration of short activities and the duration of long disturbances in a trip (Bonnetain et al., 2021).

As trajectories can describe multiple points for each trip, it is possible to infer a transport mode based on the exact locations and the measured speed between the points. This requires that the data is of high enough spatial precision and has a good temporal sampling rate (Laharotte et al., 2015). It is relatively feasible in intercity studies, where the spatial precision of the data doesn't need to be very high to be able to distinguish a trajectory that follows the train from another that follows the highway (Bachir et al., 2019). A trajectory containing an airplane trip is even easier to identify, as it contains two stops at airports separated by a conspicuously fast trip. In an urban context however, many more transport modes must be considered, such as walking, bicycling, buses, scooters, subways, or trams, which sometimes share the same geographic space, and the inference can be much more difficult (Graells-Garrido et al., 2018).

2.1.4 Link counts

Apart from flows and trajectories, another form of mobility data are **link counts**, which give the number of people passing through a link in the transportation network. The links can be segments of roads, bike lanes, or train tracks, and are usually separated by direction. Link counts are a completely different form of data than OD matrices: even for an OD matrix defined over the nodes of the transportation network, which is an optimistic situation, the

flows shown actually only say that n people were present at a first node before being present at a second node. It does not give any information about which edges of the network the people used. Even full trajectories do not give this information, as they only give a sequence of visited nodes. Conversely, link counts give information on the actual use of the network edges but no information on the zones that emit or attract the trips.

Link counts can be estimated from OD matrices by assuming that travelers always take the shortest path, or take one of a selection of shortest paths with *ad hoc* probabilities. Conversely, OD matrices can also be estimated from link counts. Most popular approaches to this assignment problem solve a formalization of it as an optimization problem (Cascetta and Nguyen, 1988; Yang et al., 2017) of the form:

$$\min_X \; F_1(X, X^*) + F_2(Y, Y^*) \quad \text{s.t.} \; \begin{cases} Y = A(X) \\ X \in \Omega \end{cases} \tag{2.1}$$

Where X^* is the prior structure of the OD matrix and Y^* is the observed link counts. The variables X and Y are the estimated OD matrix and the corresponding link counts, respectively. Ω is the feasible set for X. F_1 and F_2 are distance functions that measure how closely the output is related to the observations and the prior. As for A, it represents the trip assignment process that allows to derive the link counts from the OD matrices, that we mentioned above. Taking into account an *a priori* OD matrix allows to tackle the severe under-determination of the problem, as there are far more origin-destination pairs than the number of edges in a transportation graph. Depending on the exact form of these functions, the approaches can be separated into three categories: generalized least squares models (Cascetta et al., 2013; Vahidi and Shafahi, 2023), maximum likelihood models (Spiess, 1987), and Bayesian inference (Maher, 1983).

These models are called simultaneous OD estimations, as they require the whole data X^* and Y^* as input and therefore are restricted to offline applications. By contrast, sequential approaches solve what is called the Dynamic Demand Estimation Problem: they take as input a sequence of link count and OD matrices, and perform adjustment on the OD estimation recursively at each interval based on the current and past observations. These approaches are more suitable for real-time applications and short-term traffic prediction. The most common sequential approaches include Kalman filters (van der Zipp and Hamerslag, 1994), the least squares algorithm (Bierlaire and Crittin, 2004), and more recently deep learning (Xiong et al., 2023). A major limitation of the OD matrix estimation problem is its reliance on an *a priori* solution, which influences the result and can induce errors if it is too far from the reality (Frederix et al., 2014). To address this problem, Cantelmo et al. (2015) propose a two step approach in which the first step estimates a suitable prior before the second step estimates the OD matrix.

2.2 Data sources

The different forms of data we discussed above can be obtained from a number of sources. Each of these sources, from surveys to mobile phone tracking, has their advantages and drawbacks. Their potential depends on their precision in the various aspects described above (spatial and temporal precision, distinction by transport mode, purpose or socio-economic description), but also on the cost to collect the data and the size and biases of the population of respondents. Although it is possible to identify the potential and the limitations of any one source, it is difficult to compare them as they usually measure different objects on different populations. For example, activity chains stated by the respondent of a survey do not specify the route taken between activities, and are too limited in size to be compared with the link counts obtained on the road segments of the same region. This difficulty to compare the sources makes it difficult to quantify their limitations that we know exist, and results in the lack of ground truth for mobility data. However, the source usually taken as a reference, both because of its historical precedence and systematic efforts to control its quality, is surveys.

2.2.1 Surveys

Without containing information about how a population moves, a first source of data about the population are census surveys which give socio-economic descriptions of the population. Censuses can give aggregated statistics on the composition of the population, such as the age pyramid, but are mostly interesting in the form of **micro-samples** which individually describe each respondent. Micro-samples of population serve as a basis to model the dependencies between socio-economic variables, which can then be used to generate a population (Yameogo et al., 2020). However, they tend to be very small: the Singaporean survey HITS2012 describes around 1% of the population, while the Public Use Micro Sample (PUMS) used in Fournier et al. (2021) for the Boston Metropolitan Area covers 5%. The micro-sample made available by the French statistical institute INSEE is a notable exception in size, as it contains 30% of the population as individual rows (INSEE, 2018). It also specifies scaling factors for each individual to obtain an estimate of the total population, which are derived from the survey design.

In addition to socio-economic descriptions, the surveys usually state the home location of the respondents and can ask for their work or study locations. This allows to build a **commute matrix**, which is important information as commute trips represent a significant fraction of trips performed during a working day: in the source of activity chains that we use in this work, 31.8% of trips are commute trips. Note however that even though they share the same data structure, commute matrices are not OD matrices in the sense that they do not describe "home → work" trips. People leaving home in the morning can have an intermediate activity to do before going to work, in which case they perform instead a "home → other" followed by a "other → home" trip. In addition, a fraction of people do not go to work at all on any given workday for a variety of reasons, resulting in a constant overestimation of the number of "home → work" trips actually performed on any given day.

Household travel surveys: In addition to the socio-economic and primary location details, population micro-sample can also ask the respondents to report their full chain of activity for a given day. In this work, we call such surveys **Household Travel Surveys (HTS)**. It is

the official name of such surveys in Singapore and the USA, while they can have different name in other countries: they are called VISTA in Australia's state of Victoria, Enquêtes Ménage Déplacement (CEREMA, 2015) in France, or Deutsche Mobilitätspanel in Germany. An HTS contains a complete socio-economic description of each respondent, along with an activity chain describing the locations, purposes, start and end times of each activity, and transport modes and duration of each trip in between. It can be conducted either in person or by telephone, and a respondent may be asked to report the activity chains of the other people in their household in addition of their own. HTSs are the basis for travel demand studies and serve as the reference for most mobility indicators considered in land management (Bigi et al., 2023; Florian Schneider and Hoogendoorn, 2023). As surveys, HTSs can even include scaling factors derived from their survey design, which help to correct possible biases in the indicators.

An HTS typically contains all the information that could be useful for the analysis of mobility, with a high precision in space, time, purpose, and socio-economic description. However, they cannot be sufficient for a full understanding of the dynamics of the travel demand for the following reasons:

- First, as they are very expensive, they are only conducted every few years. In Singapore, the delay is 5 years while in France it is 10. To this must be added the processing delay, meaning that the data is already a year old when it is first available. This can be insufficient in some cases where mobility behaviors are changing rapidly, especially in a post-Covid era where working hours and locations are becoming more flexible.

- Second, the HTSs cover only a tiny fraction of the population, whereas a city can be divided into hundreds of zones depending on the desired spatial precision, for a total of billions of possible trajectories. Simply rescaling an HTS is clearly not enough to get a comprehensive overview of the location chains, or even just the OD matrix of a population.

- Third, biases are known to exist: HTSs usually suffer from a very high non-response rate, which may be correlated to the mobility behavior (Ashley et al., 2009), and people also tend to underreport the trips they have made during a past day. This is exacerbated when they report the trips of someone else in the household (Stopher et al., 2007).

These issues call for alternative mobility sources that could be used jointly with an HTS. It would be illusory to look for a replacement, as no other source could compete with the precision and the completeness of an HTS. These alternative sources should be cheap and fast to collect so that they can be used to update an aging HTS. They should also reach a significant fraction of the population so that they can be used to generalize the HTS, and contain as little bias as possible so that they can calibrate it.

2.2.2 Passive data

Passive data is data that is collected without active input from the respondent, usually through an attached device that intially serves another purpose, but sometimes through specially designed probes. This makes it both cheaper and logistically easier to collect than survey results. Unlike surveys, passive data can be collected over a specific, recent time period, and

in some cases can even be used to monitor the transportation network in real time, enabling additional applications such as short-term traffic forecasting or accident detection (Guo, 2013).

Sources of link count: Link counts on roads can be provided by a variety of traffic counters technologies, of which we mention here the most used. The earliest form of traffic counting is pneumatic tubes laid across the road, with a barometer that detects a change in pressure when a car passes (Brosnan et al., 2015). Another, more permanent method is inductive loops embedded in the road with an electric current (Oluwatobi et al., 2021). When a large metal object passes over it, it creates eddy currents that reduce its inductance, which is detected and counted as a car. Similarly, piezoelectric sensors embedded in the road can detect the pressure or the vibration exerted by a car (Li et al., 2006). Recent advances in image recognition have also made it possible to use cameras with object detection to count passing vehicles (Zhuang et al., 2009). Link count sensors often measure both the number of passing vehicles and the occupancy time of the area above them, making them able to distinguish sparse free-flows from traffic jams. They are applicable to roads but can also be applied to bicycle lanes, although as bikes are much lighter they run the risk of not being detected by inductive loops. On a rail network, the number of trains passing through a track section is generally very well controlled for operational and safety purposes, but it is not indicative of the actual number of passengers (unlike cars, where knowing the average occupancy can be enough to estimate a number of people). The occupancy of a train can be estimated using data from the braking system, which adapts to the inertia of the train and can therefore measure its weight (Bo Friis Nielsen and Filges, 2014). This can be reliable in situations where the weather does not affect the tracks, such as subways. Despite their vast variety, the sources of link counts share several common limitations:

- First, they require an investment as they rely on specific probing devices that are laid out across the network. As these devices can only measure one specific link in vast networks that can contain thousands, it can be expensive and logistically hard to implement.
- Second, they are limited to one particular transport mode, most often cars.
- Finally, the inherent limitation of link counts is that they do not measure individual coherence and thus cannot be used to estimate the factors that impact the mobility behaviors at the individual levels. This is not a problem for operational use cases and short-term predictions, but hampers the deep understanding of the dynamics of the travel demand.

Ticketing data: When it comes to counting passengers of public transport, the most straightforward way is to use the ticketing data of passengers validating their pass. Note that ticketing data can readily give passenger count data of individual trains, as discussed in the above paragraph, only in the very ideal scenario where passengers validate their pass upon entering the train, exiting it, and their pass is valid only for one specific train. This is the case for high-speed inter-city trains but usually not for regular trains and even less for subways and trams. Most subway fare systems require to tap in and out of the network, making possible the collection of OD matrices. Some fare systems such as the Paris subway, Los Angeles Metro, or typical European buses, require the travelers to validate only upon entering the network. In those cases, the ticketing data can only yield a map of trip emissions. However, it is possible

to infer the destinations in the case where the passengers have a subscription, which allows to retrieve their sequence of pass validation. By assuming that they will use the same transport mode for their next trip and not move significantly outside of the public transport network, it is possible to approximate the destination of the nth trip of any individual as the origin of their $(n+1)^{th}$ trip (Trépanier et al., 2007). It is important to note that the ticketing data does not provide information on the entire trips of the passengers, but only on the parts made in public transport. As such, the origins and destinations derived from ticketing data cannot be compared with OD matrices from other sources that would give the actual origins and destinations of the individuals. In short, the limitations of ticketing data are as follows:

- They can only measure public transport modes,
- They can only measure the trips from station to station and not the actual origins and destinations,
- Depending on the city and the exact mode, they may be unable to measure the destination station, which needs to be inferred with no guaranteed accuracy.

GPS data: The GPS systems of some cars gather location data, called **Floating Car Data**, providing to the manufacturer of either the car or the navigation system the complete trajectories of a small fleet of private vehicles (Brockfeld et al., 2007). This data is characterized by its high spatio-temporal precision, but it is restricted to cars and generally features a very limited number of people, with a penetration rate often lower than 1% (Cerqueira et al., 2018). This makes it difficult to estimate travel demand (Croce et al., 2021), but it can be used to monitor the traffic state (Kerner et al., 2005; Kyriacou et al., 2023). Alternatively, the GPS position of mobile phones can be collected by mobile applications, usually upon using the particular application in question. These sources are called Location Based Services (LBS). A good example of LBS data are Twitter records (now X) (Hawelka et al., 2014; Ahmouda et al., 2019; Porcher and Renault, 2021), which give the location of a sample of users who agreed to it each time they post on the social media. Location based services are characterized by a high spatial precision but an irregular sampling rate, as people do not use the application at short regular intervals. This severely hampers the temporal precision of the resulting trajectories. It also causes the risk of not sampling important stay points. Both Floating Car Data and Location based services data suffer from a low number of respondent, making them prone to feature a strong bias (Lenormand et al., 2014). Some private companies such as SafeGraph, Foursquare, GroundTruth, or Factual also collect trajectories from mobile apps that use their Software Development Kits (Underwood, 2019). Such trajectories feature a better sampling rate and can reach the users of multiple mobile applications. Although it could be expected, no large-scale transportation research use Google location data to our knowledge. As anyone can access their own google data, studies using it are limited to small samples of at most a few hundreds of volunteers (Hystad et al., 2022).

GPS data are trajectories, *i.e.*, collections of spatio-temporal points without semantic information that could make them readily transformable into activity chains. Extracting an activity chain from a GPS trajectory requires to interpret as activities the periods where the individual are static. However, a car can be stuck in traffic for an extended period of time, which is not an activity, but it can also make a very quick stop for an important activity, such

as dropping the children at school. The purposes of the activities are even harder to infer.

Because of their very good spatio-temporal precision, GPS traces can be used to calibrate and evaluate processing methods that generate trajectories from other data sources that do not benefit from the same precision, such as mobile phone data, which have lately been the subject of transportation research (Yang et al., 2021; Bonnetain, 2022).

Bluetooth and Wifi sensor data: Another way of collecting trajectories is to use Bluetooth or Wifi sensors laid out in a study zone (Laharotte et al., 2015; Friesen and McLeod, 2015; Naik et al., 2018). By connecting to Bluetooth- or Wifi-seeking devices, the sensors can measure the presence of the device in their coverage zone. Such devices can be mobile phones, but also modern cars that tend to be Bluetooth-enabled, making it possible to monitor traffic using Bluetooth sensors (Namaki Araghi et al., 2015). Since the devices use individual identifiers in their connections, it is then possible to reconstruct their trajectories from sensors to sensors. The main limits of these approaches come from the need to deploy physical sensors, which make it difficult to cover large areas with sufficiently small spacing. Works such as Lesani and Miranda-Moreno (2019) deploy only three sensors to monitor the campus of McGill University, while Laharotte et al. (2015) have 79 Bluetooth sensors for the whole Brisbane urban area.

Mobile data: Two sources derived from mobile phone data have been the subject of recent research for transportation applications. The first is **Call Detail Records (CDRs)**, which are operating data originally collected for billing purposes. CDRs keep track of the cellular network antenna to which the mobile phones are connected when they send a message or make a call (Wang et al., 2010). The second is **Network Signalization Data (NSD)**, sometimes called Cellular Signaling Data (Yang et al., 2021) or XDR (Graells-Garrido et al., 2018), which report all the communication events between devices that occur to operate the network. NSD keep track of the same type of information as CDRs but on a much more regular basis, as cell phones and antennas communicate to transmit technical information much more often than to transmit actual content to or from the users. Unlike CDRs, NSD are not stored by the operators by default, and they are generally only collected to investigate punctual malfunctions reported by customers. They tend to trigger several events at almost the same time, which contain different information from an exploitation perspective but no additional insight for transportation studies and make them voluminous to the point of being problematic. As a result, they can be seen as a richer version of CDRs that is more difficult for mobile operators to produce and therefore less available to researchers.

Both CDRs and NSD contain an identifier of the antenna (also called a base station), an identifier for the SIM card, and a timestamp, indicating that a given SIM card was connected to the given antenna at the given time. This can then be interpreted as the owner of the SIM card being close to the antenna at the given time, although the exact position can be hard to measure (Martínez-Durive et al., 2022). They are characterized by a low spatial precision, as the user could be anywhere within the coverage zone of the antenna with even a small probability of being outside of it, and an irregular sampling rate, which is mitigated in the case of NSD. The main advantage of cellular data is its good penetration rate, since almost everyone has a phone and there are only a handful of actual cellular providers managing the

antennas. This means that in the case of the Orange, mobile provider whose data was used in this work, about 30% of the total population is a customer and contributes to the data. As one of the main sources of data used in this work is derived from mobile phone data, we describe their characteristics and the associated processing in more detail in the next section.

2.2.3 Using mobile phone data

In both CDRs and NSD, a timestamp and the base station to which the devices are connected are logged along with an identifier. This ultimately allows to see a significant fraction of the population travelling from base stations to base stations with very little overhead costs (as the data are already logged and used for other purposes) and little delay (in the range of days instead of years for surveys). They do however suffer from a number of limitations:

- First, mobile data represent a biased sample of the population. Even though mobile operators usually reach a sizable fraction of the population, their clients bases are not shared uniformly depending on the socio-economic profiles or home locations. Correcting that bias can be difficult as the socio-economic details of the mobile users is not known in the data.

- Second, the spatial precision of mobile data is limited both by the size of the base station's coverage areas, and by the very specific working logic of base station assignment in the mobile network, which is designed to maximize the quality of service but not the simplicity of location inference (Eur, 1997, 2001). The basic logic is that the mobile phone should be connected to the antenna that it receives the best, which depends on the exact coverage map of the antennas, but also on the current network usage and the maintenance requirements. When making a phone call, the change in base station would cause a disruption in the communication, and so it is further delayed, with the result that the user can get physically close to an antenna while still being connected to the one they were at when the call was initiated. On the contrary, variations in the reception over small distances due to the complex wave propagation in a real environment mean that a same mobile phone can transfer between different antennas without moving significantly. As a result, trajectories from mobile data are known to feature oscillations between consecutive base stations (Wu et al., 2014), one-time trips to unlikely spaced base stations followed by a return to the main trajectory (Bonnetain et al., 2019), and general imprecision on the actual location of the user inside the coverage area of the antennas (Asgari et al., 2016).

- Finally, in the case of CDRs, poor temporal precision: During the day people tend to alternate heavy phone activity with little to no activity, thus disappearing from the data. Those periods are not completely at random, leading to an under-representation of people doing certain types of activities (typically, sleeping at home).

These issues of spatio-temporal precision and representativeness make mobile phone data unusable in their original state, and they require heavy preprocessing before they can be interpreted as information about the movement of a population.

Pre-processing of mobile phone data: Active research is aimed at improving this precision. If the detailed coverage map of the base station is available, it can be interpreted as

a probability of location given that a phone is connected to a base station (Forghani et al., 2020). The coverage map can can be made more precise by using the technology used for the communication (Edge, 3G, 4G, or 5G), which is associated to a specific equipment on the base station with its own coverage area. Other approaches to infer precise locations use triangulation (Wang and Chen, 2018), based on the additional information on the power of the received signal that is available in raw CDRs. These approaches are computationally expensive and require to know the coverage maps of the base stations, which are a sensitive, often updated information that mobile operators are not willing to share.

When only the locations of the base stations instead of the coverage map, research works must rely on the approximation that the users are always connected to their closest base station. The coverage area of each base station is then the set of points that are closer to it that to any other base station. The partitioning of the map in such zones is called the Voronoï diagram of a set of points. An example of such a Voronoï diagram is given in Fig. 2.5 for the antennas of telecom provider Orange in Senegal.

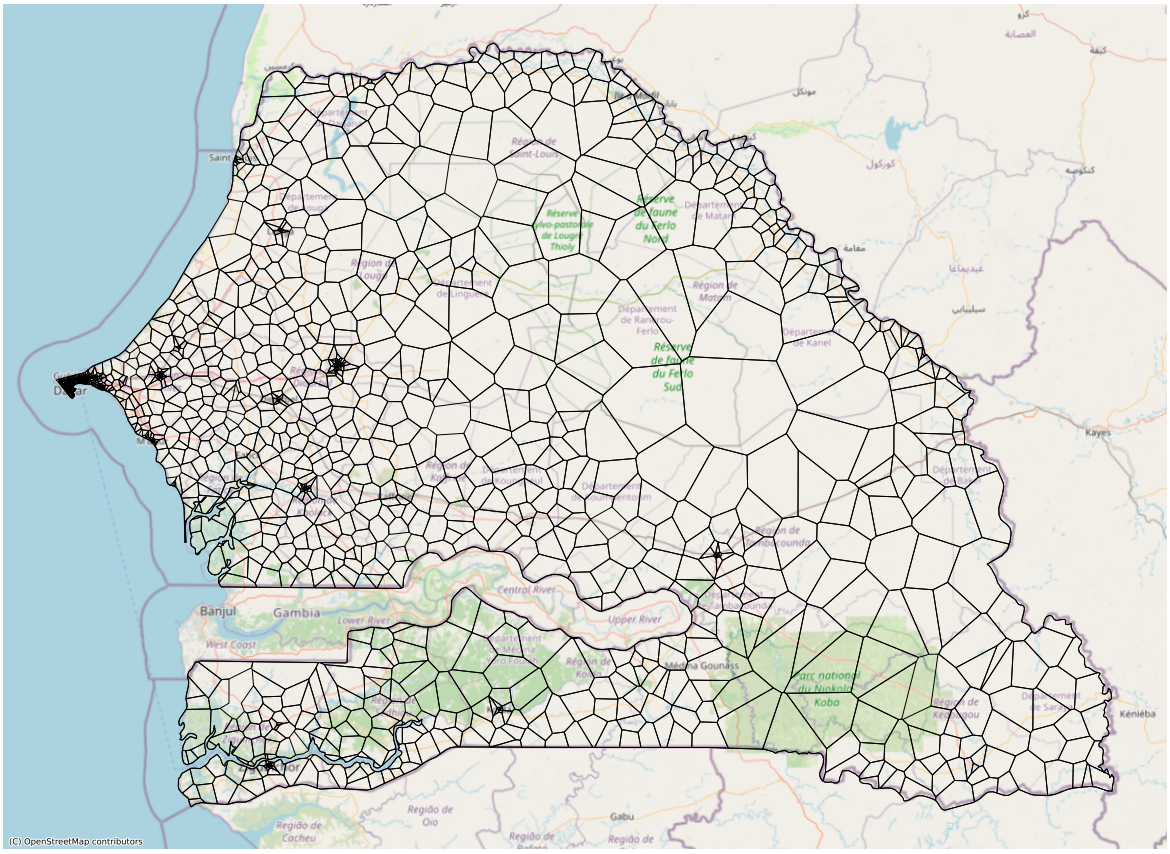


Figure 2.5: Voronoï diagram of the cellular network of Senegal of telecom provider Orange in 2013, as described in the D4D data challenge (de Montjoye et al., 2014). Raster: OpenStreetMap.

This overly simplifying assumption disregards the fact that the coverage areas are necessarily heavily overlapping to ensure a good coverage, and that the topology of the terrain has a huge influence on their shapes: a tall building is likely to stop the signal of an antenna, while concrete river banks can act as a waveguide. It also becomes impossible to differentiate

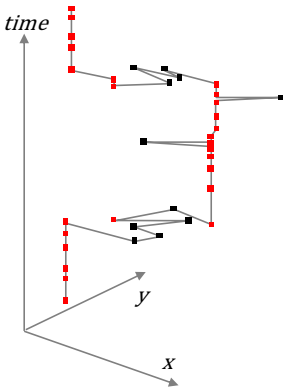
on the technology used for the communication, as all the devices are mounted on a same base station and in consequence share the same Voronoï cell. Simple adaptation such as Voronoï Boost (Martínez-Durive et al., 2022) are developed in order to improve on the Voronoï estimation without the need for additional data.

Using public data on the transport infrastructure, it is also possible to perform map matching to assign the mobile data to nodes of the transport network such as a train station or a road crossing (Bonnetain et al., 2019). These approaches assume that the mobile phone users follow a hidden sequence of nodes in a multimodal transport network, and the mobile data are noisy observations that are loosely associated to certain nodes. By formalizing the actual locations as hidden variables and the observed based stations as observed variables, the location chain becomes a Hidden Markov Model, for which the most likely sequence of location can be inferred using the Viterbi Algorithm (Asgari et al., 2016). With the same formalization, this problem can also be approached using neural networks (Shen et al., 2020). The main limitation of the map matching of mobile data is the calibration, as the models require an estimation of the transition probabilities between the nodes of the network for which no satisfying data source exist. The computing time can also be prohibitive in industrial applications (Bonnetain et al., 2019).

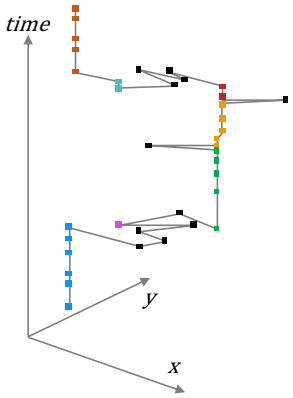
The main objective of the pre-processing of mobile data is to identify periods where the user does not actually move, which we call stay points (Bonnetain et al., 2021). The motivation is that people tend to perform activities that require staying in place for extended periods of time. The time threshold used for this detection must be carefully chosen so that meaningless stops during travels, such as waiting for a bus or being stuck in traffic, are not counted as activities, whereas short meaningful stops such as dropping children to school are still retained. A typical stay point detection would be the one used in TRANSIT (Bonnetain et al., 2021). In this processing, the antennas visited during a trajectory are first categorized between static and non static, based on whether the mobile user has spent more than a given threshold duration connected to it. Series of consecutive connections to static antennas are called static sessions. Note that a static session can contain two or more distinct antennas, as long as they were all classified as static in the first step and they are not separated by non-static antennas. Then, to account for the noise in the attribution of serving antennas, static sessions that are separated by less than a threshold number of other antennas are merged together. Finally, in an effort to filter out irrelevant stops, static sessions of less than another threshold duration are dropped.

Note that due to the low precision of the data and the heavy pre-processing, short or slow trips tend not to be detected. Indeed, the low speed of a pedestrian in urban environment can fall under the threshold of what is detected as "static", and distinct stay points that are actually covered by the same base station can appear as a single stay point.

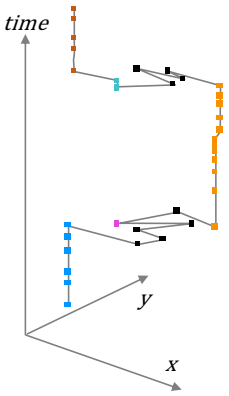
All these processing requirements must not hide the fact that trajectories derived from mobile phone data are very rich and can easily be enriched further: People tend to spend a long time at home, especially at night, which makes home detection feasible (Pappalardo et al., 2021). Many criteria can be used to elect a home antenna: the one with the most activity, the one with the most activity at night, the one that is visited in the most days, *etc.* The approaches described by Pappalardo et al. (2021) can achieve up to 69% of accuracy but



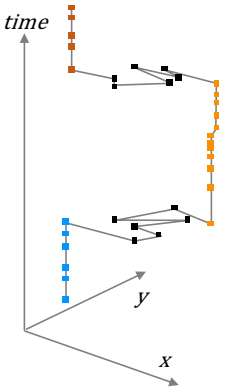
(a) Static antennas: antennas with >20min total visiting time during the day.



(b) Static sessions: series of consecutive static antennas.



(c) Merging of consecutive static sessions if they share one antenna and are separated by less than a threshold number of other antennas



(d) filtering out individual static sessions <20min

Figure 2.6: Detection of stay points with TRANSIT (Bonnetain et al., 2021).

suffers from the fact that people tend to not use their phone while they sleep. When using only CDRs, the accuracy drops to 51%.Yang et al. (2021) infer work places with a similar approach. The inference of work places benefits from using prior data in the form of the work-related Points of Interest, and by considering a more detailed representation of the antenna coverage map than the Voronoi tessellation. The authors show that the work place can be estimated with a deviation of less than 500m in 85% of the case.

Based on the transport infrastructure covered by each base station, it is also possible to infer a transport mode for the trips (Bachir et al., 2019; Zhang et al., 2022). This is especially accurate in inter-urban environment, where the mode options are limited and the antennas are organised along transport axes such as railroads or highways. This naturally gives conditional probabilities that a mobile user is traveling with a certain mode given that they connected to an antenna, from which a probability of transport mode given the whole trajectory can then be inferred.

Estimating OD matrices from mobile data: Because of their privacy liability and the computational cost of their processing, mobile data are often used only in the form of OD

matrices. The typical way of producing a set of OD matrices from a set of activity chains derived from mobile data is to consider a temporal partitioning, usually by time steps of 15 minutes to 2 hours, and a spatial partitioning, and report only the number of trips performed for each tuple of origin o, destination d, and time step t (Iqbal et al., 2014). Considering both the social benefit of understanding the movements of a population and the economic opportunities, various mobile phone operators now propose commercial solutions to sell OD matrices derived from their CDRs or their NSD. These include solutions such as Telefonica NEXT in Germany, or Orange Flux Vision in France.

2.3 Mobility models

The data sources described above are a valuable input to understand the dynamics of a transportation network, but a full understanding can only come through the modelling of the data. This is true in science in general, but in transportation research the modelling carries extra value, as it can then help to generate mobility data. Generating mobility data can help solve three main problems:

- Scaling the data to the whole population: the richest data sources tend to have a very low number of respondents, making them prone to a high variance and possible bias. Statistically, this small sample size combined with the high dimensionality of spatial data makes it impossible to simply apply scaling factors to data such as trajectories. In order to capture the high diversity of mobility behaviors from small samples, it is important to be able to infer behaviors that have not been observed, which can only be done by a careful modelling of the individual choices.

- Integrating several sources: all the sources presented above have their shortcomings, either in terms of representativeness or completeness of the data. Generating data offers the opportunity to calibrate it on different aspects by different data sources, allowing to use the one considered the most reliable for each aspect.

- Estimating the data in an hypothetical scenario: describing the current state of the network is helpful to guide public policies, but it can be even more valuable to predict their impact. This requires to be able to generate data that depend on a context that can be edited. Analyzing the response of the network to different hypothetical situations also allows to better grasp its sensitivity to various factors, opening a new path toward actually understanding the network. This is of particular interest as the response of transportation networks is known to be counter-intuitive at times, with numerous documented situations of traffic evaporation where closing roads have led to an overall diminution of the traffic in the area, with no clear understanding of where it went (Cairns et al., 2002).

Mobility models can be separated into two categories: 4-steps models focusing on aggregate flows (we say that they are trip-based), and activity-based models relying on individual choices.

2.3.1 4-steps models

These models developed in the 60's have been the historical solution to produce transportation forecast (Bonsall, 1997). They actually consist of four separate, mostly independent models, each generating new information that improves on the previous output.

- Trip generation: The first step produces estimations of trip emission and attraction for each point of interest in the study zone. In other words, it estimates the number of trips generated from, and attracted to, specific locations within the area. It can be inferred using proxy variables related to the land usage, such as the number of households or the number of employees at the location.

- Trip distribution: The second step, based on the emitted and received volumes, constructs an OD matrix representing the flow of trips between different origins and destinations within the study zone. A popular model for this is the gravity model (Isard, 1954), stating that the volume $v_{o\to d}$ of a flow $o \to d$ increases with the emission of the origin o noted E_o, the attractiveness of the destination d noted A_d, and decreases with the distance from o to d noted D_{od}:

$$v_{o\to d} = K\frac{E_o^\alpha A_d^\beta}{D_{od}^\gamma},$$

  where α, β and γ are hyper-parameters that can be tuned to fit potential OD matrix data. As an alternative, the radiation model proposed by Simini et al. (2012) accounts for the competitive opportunities inside the circle of radius D_{od}. Separately from 4-step models, both methods are adapted to model general trends such as commute patterns (Stefanouli and Polyzos, 2017), migration flows (Alis et al., 2021), or trade flows (Kabir et al., 2017). Note that with these approaches, the trip emission and attraction inferred in the first step is not interpreted as an actual number of trip that will go in and out of the zone but more as an indicative score.

- Mode split: The third step categorizes the trips based on the transportation method used, such as car, bus, or subway. This is generally done via discrete choice models (Vernydub, 2020).

- Trip assignment: The fourth step assigns the actual route taken to go from an origin o to a destination d. This step determines the specific paths that travelers would take based on the available options and the capacity of the transportation network. This step is generally formalized as the maximization of individual utilities in which we solve for user equilibrium (Beckmann et al., 1955).

Such models are adequate for highly aggregated analysis for strategic predictions, and are for example still used for long term forecasting of autonomous vehicle adoptions due to their low demand on input data (Dias et al., 2020). They have been criticized however for being unable to capture the interactions that drive each single choice. They only consider total flows and assign each new piece of information independently from each other (Rasouli and Timmermans, 2013). In other words, these models may not effectively account for the complexities of individual decision-making and the dynamic interactions between different components of the transportation system.

2.3.2 Activity-based models

With the increase of available computational power, mobility studies have begun to rely on much more detailed models that simulate individuals instead of forecasting aggregated flows. These new models are called **activity-based models** as they rely on the consideration that the mobility we observe are due to individuals moving with a particular purpose (Arentze et al., 2000; Galli et al., 2009; Hilgert et al., 2017). Understanding the way people choose their activities, their locations, and the specifics of their trips is then the key to understand the mobility of the population.

Activity-based models propose to generate a population of agents with their individual features, constraints, and rules for finding an activity chain matching their constraints. They do not handle trip assignment, which is usually left to a transport simulator such as MATSim (Horni et al., 2016), SimMobility (Adnan et al., 2016) or POLARIS (Auld et al., 2015). In that aspect, activity-based models can be seen as handling the steps of OD matrix generation and transport mode assignment of the traditional 4-steps model, while the transport simulator handles the trip assignment. The main difference would be that instead of considering rules over aggregated flows, activity-based models and transport simulators consider individual agents. The various aspects of the transport demand synthesis have separately been the object of research:

- Generating the socio-economic features of the agents. This problem, known as the **synthetic population** problem, is to generate agents defined by their socio-economic features, such that the resulting synthetic population resembles the true one. More formally, we can interpret each individual as an independent realisation of a multivariate distribution, which has to be estimated. This problem amounts to generating a micro-sample of the population of the desired size, generally the whole population or a significant fraction of it. It can be done by expanding a small pre-existing population sample (Beckman et al., 1996; Ye et al., 2009; Bar-Gera et al., 2009; Müller and Axhausen, 2010; Fournier et al., 2021), or more recently by modelling the joint distribution of variables and sampling from it (Sun and Erath, 2015; Ilahi and Axhausen, 2019; Borysov et al., 2019; Garrido et al., 2020)

- Assigning activity chains to the agents. Activity chain assignment is the problem of endowing each agent, at this point only defined by socio-economic factors, with a chain of activities defined by their purposes, time of the day, and optionally the transport mode between each activity. This is done either by designing a generative process (Liu et al., 2015; Saadi et al., 2016; Joubert and De Waal, 2020), or by assigning existing activity chains to the agents of the synthetic population (Hörl and Balac, 2021; Namazi-Rad et al., 2017).

- Setting locations for the different types of activities. This problem is generally separated in two steps, as primary locations such as home, work place, or study place can be inferred from commute matrices (Hörl and Balac, 2021), population micro samples, or approaches such as radiation or gravity models described above. Secondary locations are much harder to infer, mostly because of the high number of possibilities and the little data available. They are usually sampled inside a restriction of the space called a **space-time prism** (Lenntorp, 1976; Yoon et al., 2012; Justen et al., 2013), that describes

the locations attainable between two primary locations given the available time, with a maximum travel speed. More recently, mobile phone data has been a promising source to draw such location (Ballis and Dimitriou, 2020; Anda et al., 2021)

These activity-based models have nothing in common with 4 steps models in term of resources required: simulating millions of interacting agents take hours on modern computers, and would have been unfeasible in the 80's. But the main difference is the amount of data required to calibrate an activity-based model, as it needs an individual description of all the persons in a population with their mobility behaviors, on top of a comprehensive map of the infrastructures and estimations of a large number of flows.

2.3.3 Data requirements

4-steps models have been designed to use the data that have been historically available, that is surveys and geographical information. This rather coarse, static data mostly focuses on aggregate statistics of the population and characteristics of geographic zones, such as the number of homes or work opportunities, the distances that separate them, or the transport infrastructure. It is often available and its processing is rarely a problem.

By contrast, the focus of activity-based models on people's organizations, choices, and motivations requires detailed data that captures the structure of individual activity chains. In fact, activity-based models can be seen as addressing the problem of generating a synthetic HTS that would cover an entire population. Doing so requires data on the socio-economic composition of the population, usually in the form of a micro-sample to capture the dependencies of the variables, and aggregated statistics to ensure the correctness of the resulting synthetic population. It also requires primary location data, which can be obtained by commute matrices estimated from surveys. The structure of activity chains can be obtained from an HTS. However, surveys cannot reach the exhaustivity required to appropriately estimate the complete chains of locations, an attribute that can have billions of modalities.

Mobile phone data comes as the last piece for comprehensive activity-based models. In the last fifteen years, what was initially operational records kept only for billing purposes without any attention given to the precise locations became a major source of insight into the mobility behavior of the populations. CDR data works especially well for the monitoring of long travels between large zones of interest, where we can consider the size of the coverage area of the base station to be negligible. At smaller scale, typically cities, their low spatial precision and their noisy behavior make them less than ideal for transportation studies. In fact, it is widely admitted now that mobile data alone are a sufficient source to produce mobility reports on large-scale geographies, but that attempting to do the same on a city would yield at best imprecise, at worst misleading indicators on the mobility of the inhabitants (Wu et al., 2014). Where it is customary for activity-based models to give information at the address level, mobile data can hardly do better than indicating a base station that typically covers a region of hundreds of meters of radius, but whose actual geometry is generally unknown from researchers. Apart from the underground subway which is equipped with its own dedicated base stations, it is mostly impossible to infer a transport mode from a mobile phone trajectory in a city where pedestrians, bikes, cars, buses, trams, and all kind of novel vehicles share almost the same space. As for the rescaling of the data, it is generally based on the home

location of the mobile users but it is very optimistic to suppose that the penetration rate of the operator does not depend on other, socio-demographic factors. These limitations make it necessary to use mobile phone data jointly with other sources that can handle the individual details. An activity-based model could combine mobile phone data with an HTS to exploit the strengths of both sources.

2.4 Respecting privacy

A natural concern when handling transportation data is their privacy risk. Transportation data contain information that we call personal for two reasons. First, because they can describe the whereabouts of one particular individual. Second, because they give information about the behavior of groups of individuals that can be more precise that what the individuals would deem comfortable. Fortunately, transportation studies are not interested in localizing one particular person, and the main challenges in transportation are when the mobility demand is high. This means that the individual details of the travelers are never of interest in themselves. They are however collected in the form of individual trajectories or micro-sample of population before being converted into operational indicators. It is then important to make sure that these data sources do not present a privacy threat, while keeping the useful information about the behavior of the system as a whole.

The very concrete ways in which a data set can divulge personal information depend on who asks for it. Classically, the literature considers an ill-motivated data user that we call an attacker, who wants to use the data set to know more about one particular individual, that we call the target (Fung et al., 2010). The attacker may already know something about their target, which we call the prior information, that they acquired either because it is public knowledge (like the basic information on a celebrity), because they know the target personally, or because they accessed other data sets that may or may not contain personal information.

The intents of the attacker can also vary. We classify the different types of privacy attacks depending on their scope and what is considered successful for the attacker (Gramaglia and Fiore, 2015):

- A **record linkage attack** aims at uniquely identifying the target in the data set. In case of success, the attacker can access all the information that the data set contains about their target. From an attacker's perspective, this is the most difficult type of attack.
- An **attribute linkage attack** aims at acquiring an exact information about the target, even if the target in itself is not pinpointed in the data. A typical example in health data would be a hospital record where all 59-year-old men have the same type of cancer. In that case, an attacker who wants to know more about someone who happens to be a 59-year-old man can infer his health status without having to pinpoint him in the data.
- A **probabilistic attack** aims at updating the prior knowledge of the attacker using the data. This is the easiest kind of attack from an attacker's perspective, since a situation like the one described above but where 59-year-old men have different pathologies can still be a success for a probabilistic attack. It is enough that the distribution of the values inside the group is different from what the attacker could guess before accessing the data. Note that it also means that probabilistic attacks are the ones carrying the

smallest stakes, as the knowledge update can be almost insignificant.

It can be difficult to protect a data set against all possibilities of attacks, especially since the attribute linkage and probabilistic attacks can technically be successful without the target being actually in the data, as long as we consider that the data set is representative of a population to which the target belongs. The typical example is a study over whether smoking cause cancer (Dwork, 2011). The details of the respondents of such a study would clearly be personal information and shouldn't be publicly released. But the conclusion in itself, that smoking does cause cancer, can also be interpreted as personal information as it means that knowing that someone smokes gives an attacker additional information on their general health.

It is then worth discussing what conclusions constitute a valuable insight on the health or the behavior of a population as a whole, and what constitutes a privacy breach. The question is easily answered in the case of medical studies since lives can be at stake. But in the case of transportation, the conclusions are not of absolutely vital interest and some conclusions may simply not be worth the privacy risk. If an OD matrix shows that every single person who exits the subway at one particular station goes to the nearby hospital, then the privacy of the people who go out of this subway station is at risk as anyone knows that they either have health issues or one of their relatives has, all this for limited gains from a point of view of transportation analysis.

This idea that safe data should not give out additional information on particular people regardless of the prior knowledge of the attacker is called **Uninformativeness** (Machanavajjhala et al., 2007). That a data set stricly respect uninformativeness is a very strong and arguably sufficient criterion to deem it safe regarding privacy. However it is also very optimistic to expect all data sets to guarantee uninformativeness, since the information given by any one particular data set can contribute to the attacker's prior for a better targeted attack on any other data set. With this view, even a subway map can appear as a privacy liability, as an attacker observing their target being in the subway at the next-to-last station of the line can use the subway map to infer the exact station where the target gets off. A looser criterion is **indistinguishability**, which is considered satisfied if attackers cannot pinpoint any individual in the data. The strictest implementation of the indistinguishability principle is the property of k-anonymity, which holds if in a data set, any individual row is exactly equivalent to at least $k-1$ other rows. Indistinguishability does not guarantee safety against attribute linkage and probabilistic attacks. For example, the target can belong to an outlier group, very distinct from the others, in which case attribute linkage attacks will be successful. The motivation behind indistinguishability is that if the information concerns a large enough group of people, then the public is justified in knowing about it.

2.4.1 Pseudonymization is not enough

It is a common mistake to expect that data that does not contain any name, e-mail address, or phone number would be anonymous based on the assumption that an attacker couldn't retrieve one particular person. Suppressing such **direct identifiers** and keeping only a pseudo-identifier to maintain the individual consistency is known as **pseudonymization**. While it does make it harder for an attacker to retrieve their target, it offers no guarantee to make it strictly impossible. The combination of all pieces of information about the

pseudonymized individuals can be enough to identify someone in the data. A famous illustration of this problem is given by Sweeney (2002b) using Cambridge Massachusetts voter's list and Massachusetts hospital records, which at the time had been been public and available to anyone for a 20 dollars fee, respectively. The voter's list contained the name, ZIP code, birth date and sex of all the registered voters, whereas the medical data contained the ZIP code, birth data, sex, and medical condition of 135,000 state employees and their families. It was believed to be anonymous because the name was removed, which is equivalent to pseudonymization in this case since all the information about any one particular individual were contained in one row of the data. However, then governor of Massachusetts William Weld was the only one with his combination of ZIP code, birth date and sex in the medical data, meaning that the exact reason of his visit to the hospital were actually public.

This problem is also present in the case of a location chain, made of spatio-temporal points denoting activities made by an individual. The natural identifier for the chain, in the case of mobile data, can be either the SIM card identifier or the device ID uniquely identifying it. Pseudonymizing the trajectory consists in replacing this identifier with another, random identifier. Yet the trajectory in itself likely contains enough information to re-identify the individual, most of all because the trajectory is likely to state the home and the work locations. If it is of sufficient precision, there is a good probability that only one person lives and works at two given arbitrary locations. Knowing only these two locations, an attacker could then learn the whole trajectory of their target.

A famous study by Montjoye et al. (2013), based on a data set of 1.5M mobile phone user trajectories over 15 months, showed that by drawing just 4 random points of any CDR trajectory, in 95% of the cases this trajectory was actually the only one containing all these points and thus could be uniquely identified. In practice, the prior knowledge an attacker could have about their target is not completely random. Another study by Zang and Bolot (2011) shows that in a data set of 25M mobile phone users, more than 50% of them can be uniquely identified by their 3 most visited locations. This can be problematic as the most visited locations of an individual are precisely the kind of information of interest we could want to keep based on mobile data.

Note also that a desirable property of anonymization, which happens to be a legal requirement in European legislation, is that it should be irreversible regardless of the other data set that an attacker could access. In particular, if the mapping table from identifiers to pseudonyms can be retrieved by any means, then the pseudonymization offers no protection at all.

2.4.2 Anonymization of location chains

In fact, making a set of location chains meet a privacy criterion can only be done at the expense of the general quality of the data, as it requires reducing either the precision of the data, its volume, or both. Trajectories are known for their high variety, with a high number of possible values and numerous outliers (Gramaglia and Fiore, 2015). In this context, it is impossible to guarantee k-anonymity for $k > 3$ without destroying all utility of the data. Although this is technically enough to guarantee safety against record linkage attacks, it still constitutes an arguably weak privacy guarantee.

This has lead research to consider weaker or alternative forms of indistinguishability, that rely on assumptions over the prior knowledge available to the attacker and ensures indistinguishability only in those scenarios, at the risk of failing if an attacker as more knowledge that what was assumed. Such definitions include k^m-anonymity (Terrovitis et al., 2008), which states that any individual should be indistinguishable from at least $k-1$ other individuals as long as the attacker only has access to at most m attributes. For a data set of d attributes, k^m-anonymity is equivalent to k-anonymity if $m = d$. For other values of $m < d$, then k^m-anonymity is a relaxation of k-anonymity in the sense that a given individual does not have to be hidden in the same exact group for all features. A small example of a k^m-anonymous is given in Table 2.3, where each individual is distinct from all the others, yet if an attacker tries to identify their target using only two attributes there will always be at least two corresponding rows.

	age	**city**	**gender**	**car**
1	25	Paris	female	no
2	25	Paris	male	no
3	25	Paris	female	yes
4	25	Paris	male	yes
5	25	Lyon	female	yes
6	25	Lyon	male	yes
7	42	Lyon	female	no
8	42	Lyon	male	no

Table 2.3: Example of a k^m-anonymous data set with four variables, with $k = m = 2$.

The logic of k^m-anonymity has been adapted to trajectories by Gramaglia et al. (2017), where an attacker is supposed to only access the points of the trajectory that fit into a time window of duration τ. However, this relaxation of the k-anonymity is associated to an additional restriction against attribute linkage attacks: the attacker should not be able to perform a successful attribute linkage attack on more than another temporal window of duration ϵ, disjoint from the first. This means that when ϵ is small, it is not only a possibility but a requirement that any one particular trajectory be hidden alongside some trajectory at the beginning and some other trajectory at the end. An illustration of $k^{\tau,\epsilon}$-anonymity is given is Fig. 2.7: trajectory i, defined by its spatio-temporal points, is indistinguishable from at least one other trajectory as long as the attacker can only look at points in a time window of duration τ. But since trajectory c is the only one to follow trajectory i for an extended period of time, knowing the particular points in the window illustrated in the figure implies that the attacker knows where the target during all the illustrated time window of duration ϵ. As this is not true for a window longer than ϵ, then ϵ is the upper bound of the privacy leakage.

2.4.3 Anonymization of other mobility data

Achieving k-anonymity with higher values of k is possible for the other forms of mobility data that are OD matrices, link counts, and emission-reception maps. In the case of the two

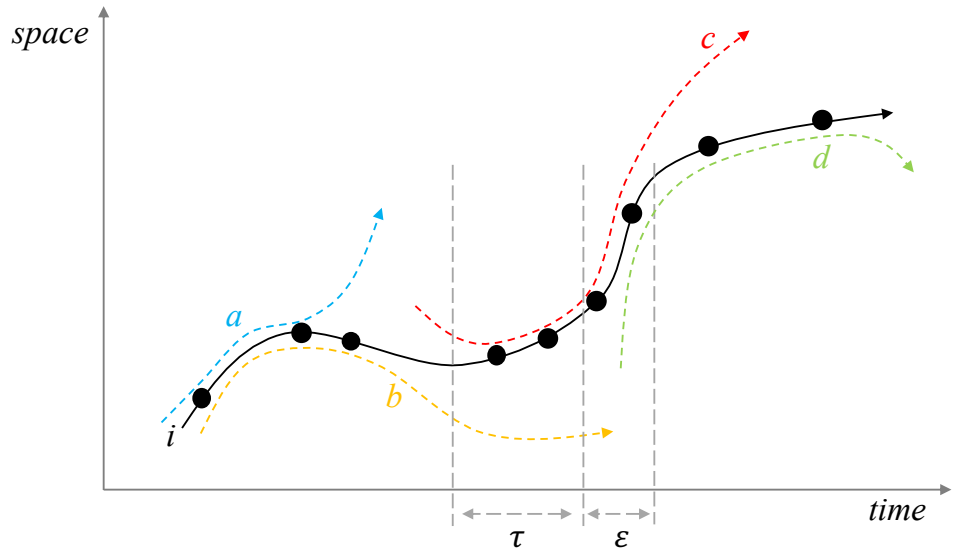


Figure 2.7: A data set of trajectories that is $k^{\tau,\epsilon}$-anonymous for trajectory i. Example inspired from Gramaglia et al. (2017).

latter forms, it is enough to consider a coarse time precision to ensure that no reported count is below k. In fact, even without k-anonymity, it can be argued that appearing in the data as the only person to have passed by a road link at a given time would hardly constitute a privacy risk. It would indeed be the only available information in the data set, and an attacker would have to know it already in order to retrieve the person. In this case, k-anonymity is rather an additional precaution ensuring that even in the most unlikely circumstances, re-identification is impossible.

In the case of OD matrices, the remark is still valid to some extent but the circumstances where an OD matrix can disclose personal information can be more common. Knowing that the target leaves a given area at a given time can easily lead to only one or two possibilities of destination in a time-dependent OD matrix, as long as the spatial and temporal precision are high enough to make the matrix highly sparse. OD matrices can be seen as set of location chains that contain only two points, and similarly to location chains they maintain their high amount of outliers and high number of values. It is then necessary to ensure that these outliers are not uniquely identifiable in OD matrices, which is usually done by aggregating the zones of the OD matrices. Chapter 3 develops our approach to the problem of finding an aggregation that minimizes the loss of information while satisfying the constraint of anonymization.

2.5 Conclusion

This review of the context of travel demand illustrates the variety of sources that are available, each characterized by the object they describe, their level of details with respect to different aspects, and their biases and shortcomings due to to their nature. Link counts can give a reliable estimate of the demand on one particular road segment, but cannot give information on the origins and destinations, and even less at the individual level. Ticketing data gives the origin but can struggle to give the destination, and in any case give only the origins and destinations inside the public network, and not the actual goal of the user. Surveys

and GPS data have in common that they tend to reach only a handful of people, while mobile phone data suffers from poor precision. Yet they all give an insight into one aspect of the true behavior of the city's transport network.

The value of these sources is harnessed in a variety of mobility models in an attempt to summarize the mobility, and to explain it. Recent advances in activity-based models, which attempt to model the motivations and constraints of the individuals, require large scale, detailed data sets in order to maximize their realism. As each data source tends to be reliable on one aspect but less so on the others, this contexts calls for ways of merging several data sets that are *a priori* very distinct from each other.

Finally, it appears that most of the latest sources, which have the potential of giving large-scale spatial information on people's mobility, tend to be hard to come by not for logistic reasons but rather because of privacy concerns. Anonymizing such data then appears as paramount for the development of this research field.

Chapter 3

Anonymisation of OD matrices

3.1 What is anonymization and why it is difficult

Transportation studies often suffer from a lack of data, either because the data is of poor quality, the sample is too small to draw conclusions, or is simply unavailable.This is partly due to the logistical difficulties of collecting mobility data, as it requires probing a large number of people for a lot of information, but the main problem is its personal aspect: even though mobility data does not contain directly identifiable information such as name, address, date of birth, or social security number, as can be the case with health data, people are still generaly uncomfortable sharing their whereabouts. This is quite natural, as someone's exact trajectory is usually unique given the high number of degrees of freedom (recall the cases studied by Montjoye et al. (2013) and Zang and Bolot (2011), mentioned in Sec. 2.4).

Recently, several strong restrictions on the use of personal information have been enacted around the world. Laws such as the Personal Data Protection Act (PDPA) in Singapore, the California Consumer Privacy Act (CCPA), or the Personal Data Protection Law (PDPL) in Argentina all aim to define what personal data is, under what conditions it can be used, and how it can be made anonymous. The European General Data Protection Regulation (GDPR) is known for being particularly restrictive and influential among them (European Commission, 2018). Its influence comes from the fact that any company targeting European residents must abide by it, making it simpler for foreign companies to apply it to all of their customers, which in turn encourages foreign legislation to enforce similar laws (Roberts, 2018). This welcome legislative concern for privacy acts as a strong disincentive for all stakeholders to share their data, hampering efforts in transportation research.

However, the personal aspect of the data is not a necessary condition for insightful transportation studies: the constraints of modern transportation networks arise from the large number of their users, requiring high capacity networks that can handle flows in which the presence of any particular person is irrelevant. The models, reports, and conclusions that researchers and administrators produce and use are all designed to identify trends or to report aggregate statistics, and have no interest in detailing specific behaviors. On the contrary, their interest is to identify societal trends that may have an impact on public arrangements, with the goal of organizing a greener, more efficient, and more resilient transportation net-

work. This means that there is a possible compatibility between high quality and anonymity in travel data. The problem then becomes to identify the limit between insightful descriptions of a social trend and personal information that threatens privacy.

3.1.1 The various definitions of anonymous

As mentioned in Sec. 2.4, pseudonymization is never considered as providing sufficient anonymization guarantees under either the GDPR nor state-of-the-art (Sweeney, 2002a). The privacy liability of a data set is measured as its vulnerability to the three kinds of attacks:

- Record linkage attacks, when an attacker tries to identify their target in the data set;
- Attribute linkage attack, when an attacker infers the value of a particular variable of their target;
- Probabilistic attacks, when the attacker updates their prior information in any way about the target.

Each one of these three attacks is a relaxation of the previous one, where the goal is less ambitious for the attacker but the success is more likely. The potential for success of each of these attacks depends on the quality and quantity of prior knowledge available to the attacker. The state-of-the-art in this regard usually distinguishes between **sensitive information**, which is never known from the attacker *a priori* and is the targeted information, and the **quasi-identifiers**, which are considered harmless in themselves but can help identify someone in the data. A typical attack scenario would be to assume that the attacker knows exactly the values of the all quasi-identifiers for their target, and wants to update their knowledge on the sensitive attribute in the data.

However, under the GDPR terminology, any attribute related to a specific individual is **personal information** and should be protected. The concept of sensitive information under GDPR refers to certain personal data such as biometric data, racial and ethnic origin, political opinions, religious or ideological convictions, and their collection, storage and processing is basically out of the question. In this sense, the state-of-the-art approaches must be adapted to consider all quasi-identifiers as sensitive (in the sense of the state-of-the-art, *i.e.,* personal in the sense of the GDPR). Throughout this work, we will only use the GDPR terminology.

The most difficult protection against privacy attacks is then to assume that an attacker has perfect knowledge of all but one attribute of the data, and wants to update their information about the missing attribute. Assuming this attack scenario allows for **Privacy-Preserving Data Publishing (PPDP)** (Fung et al., 2010), which is a principle designed to guide the best practices in data anonymization. The two main principles of PPDP are:

- Publication of data, not of data mining results: Anonymization should not assume any particular subsequent use of the data. This means that all attributes should be preserved as much as possible, which is much more restrictive than assuming that some information will not be useful and discarding it.
- Truthfulness at the record level: Each data point should correspond to an actual record in the ground truth. This prevents us from creating fake individuals that would flood the data, or applying noise or permutations in the data.

The most important anonymization criterion following PPDP principles would be ***k*-anonymity** (Sweeney, 2002a). It is a formal privacy guarantee used both in research and by authorities such as the French regulator CNIL. A data set is said to be k-anonymous if each individual in the data is indistinguishable from $k-1$ other individuals. Technically, k-anonymity with k as low as 2 is secure against record linkage attacks, although this would qualify as a rather weak anonymity condition. In particular, it would provide very limited protection against attribute linkage and probabilistic attacks. The state of the art usually aims for k between 3 and 5 for hard to anonymize data sets, and up to 200 for simple data sets (Liang and Samavi, 2020).

A complementary property is ***l*-diversity** (Machanavajjhala et al., 2007), which holds when each generalized value covers at least l distinct modalities. Together, k-anonymity combined with l-diversity ensure protection against attribute linkage attacks, and significantly reduce the potential of probabilistic attacks (Shokri, 2015).

For even further protection, it is possible to implement ***t*-closeness** (Li et al., 2007), which holds when the distribution of attributes in each group of k is no further than a threshold t from the total distribution. This threshold can be interpreted in terms of any probability distribution metric, such as Kullback-Leibler divergence or the Earth mover distance.

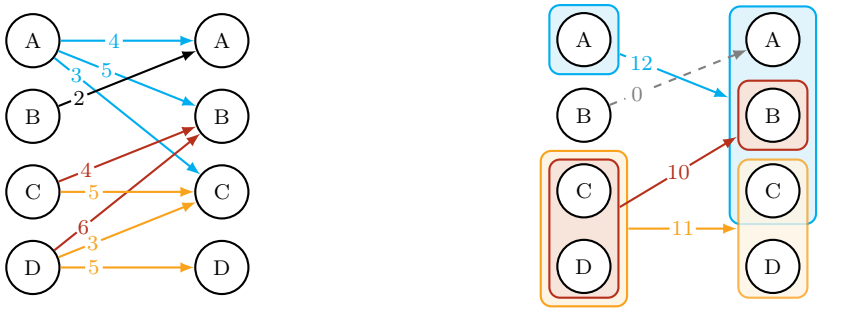
For each of these three criteria, the corresponding threshold acts as a parameter that can be tuned to enforce more privacy at the cost of utility of the data, or conversely be more lenient at the risk of putting privacy at risk. However for OD matrices, it can be argued that k-anonymity alone, given a sufficiently high k, is enough to make the individuals no longer identifiable: the areas of origin and destination can be large enough to avoid attribute linkage, and the small number of attributes makes probabilistic attacks rather vacuous.

Note that the relaxations of k-anonymity mentioned in Sec. 2.4, such as k^m-anonymity (Terrovitis et al., 2008) and $k^{\tau,\epsilon}$-anonymity for trajectories (Gramaglia et al., 2017), either assume limited prior knowledge on the part of the attacker or accept some form of data leakage. As such, these adapted criteria are not compatible with GDPR.

3.1.2 Anonymization as a combinatorial optimization problem

The main approach to k-anonymization is **generalization and suppression** (Sweeney, 2002a), *i.e.,* reducing the precision of the data to form indistinguishable groups of at least k individuals and suppressing the individuals that are not in such groups. In the case of OD matrices, this means that the origin and the destination of flows are made coarser so that they merge until their total volume reaches more than k. A toy example of generalization and suppression for the 10-anonymization of an OD matrix is illustrated in Fig. 3.1: flows $A \to A$, $A \to B$ and $A \to C$ are aggregated together in order to reach a volume above 10, and similar aggregations are performed on the flows originating from C and D, and directed towards destinations B, C and D. Flow $B \to A$ is suppressed, which can be preferable if hiding it in the other flows requires large generalizations.

Finding such a solution that minimizes the loss in precision is known to be a NP-hard problem (Bettini et al., 2005). Historically, the first k-anonymization algorithm was Datafly (Sweeney, 2002a), which relies on a generalization hierarchy describing how modalities should be merged



(a) Initial OD matrix

(b) 10-anonymous OD matrix

Figure 3.1: Example of generalization and suppression of a simple OD matrix. Note that the flows here have been clustered without any kind of constraint. Some approaches seek particular solutions, notably ones where the generalizations form a partitioning of the domain.

together. An example of such a generalization hierarchy is illustrated in Fig. 3.2: for a data set giving the position of individuals in a study zone, the initial modalities are represented by the leaves of the tree. If we choose to generalize the whole data set one level higher, then the possible modalities are the parents of the leaves. Datafly finds the best horizontal cut in the hierarchy, meaning everyone in the data is generalized to the same level. This **uniform generalization** approach has the advantage of being scalable to huge volumes of data, as well as data with numerous attributes. In a bid to find a finer-grained result, some approaches look for a generalization at the individual level, which gives a solution akin to a clustering (Liang and Samavi, 2020). In the specific field of mobility data, this approach is better represented by Glove (Gramaglia and Fiore, 2015), which generalizes points in a data set of trajectories. However, these approaches lack scalability for data sets characterized by a huge number of flows, as is the case for data derived from mobile phone CDRs and NSD: for example, Liang and Samavi (2020) report computing times in the order of hours for low values of k, and their approach can be expected to take days for $k \geq 10$ given its exponential time dependency. In order to achieve fast, accurate anonymization of OD matrices, dedicated processing is needed.

In this chapter, we propose a methodology to efficiently make OD matrices comply with the restrictive definition of anonymous data in the GDPR, *i.e.*, where the data subject is not or no longer identifiable (Recital 26 GDPR). According to this definition, European regulators consider small flows to be personal data, making the distribution of OD matrices in the European Union problematic, even though they are admittedly much more secure than the trajectory collections from which they are derived. Since k-anonymity is recognized by the French regulator CNIL as an acceptable anonymization measure, it is an appropriate criterion to implement in order to promote the dissemination and use of OD matrices.

Note that like mobility data, OD matrices are characterized by their uniqueness: the flows can be small and isolated from others, and the origins and destinations can be among sets of up to thousands of areas. Compared to regular, relational data sets, this high number of possible values make OD matrices more difficult to anonymize. Moreover, mobile phone data are characterized by their huge volumes, expressed in number of distinct flows. They also have a high velocity, meaning that they are generated at a high rate and so must be processed regularly. This calls for an anonymization method with a low computational cost that can

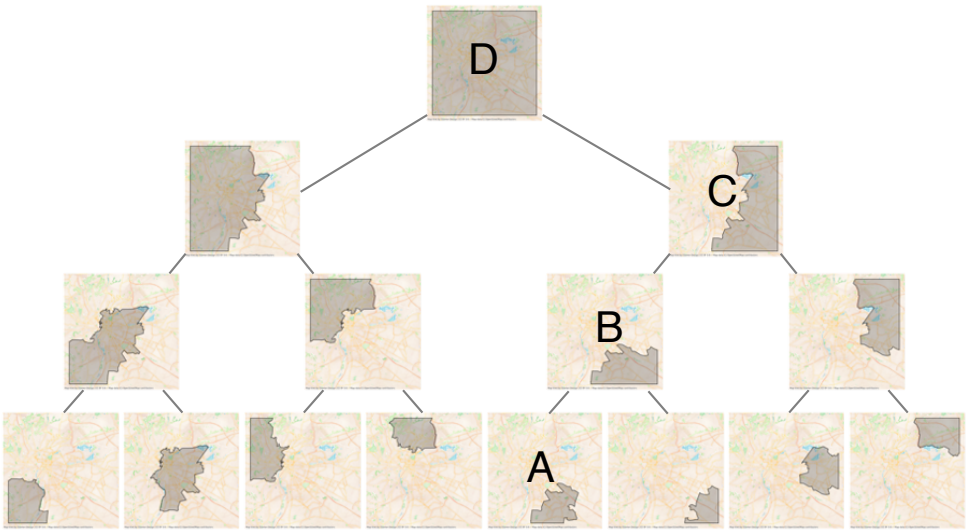


Figure 3.2: Example of a spatial generalization hierarchy. The root represents the whole study map, and the children of a node form a partitioning of the parents. An individual present in area A in the data can be generalized to be shown as present in area B, C or D depending on what is necessary in order to hide them in a group of k individuals.

quickly process big OD matrices.

In this chapter, we develop an approach for k-anonymization of OD matrices with a focus on scalability. Our contribution is three-fold:

- We formulate generalization and suppression over a generalization hierarchy as an optimization problem, taking advantage of the low dimensionality of OD matrices in order to broaden the search domain of solutions. As representativity is paramount for the value of mobility data, we also control the volume of suppressed records by keeping it under a fixed constraint. We solve it by using state-of-the-art algorithms for dependency-constraint knapsack problems, finding an **adaptative generalization** finer than uniform ones. Based on this formulation, we propose Adaptative Tree generalization (ATG), a lightweight algorithm to efficiently reach k-anonymity. The algorithm is proposed in two versions, **ATG-Dual** and **ATG-Soft**, corresponding to two variations of the problem we solve.

- We evaluate our approaches against an extensive benchmark of anonymization methods from the state of the art: uniform generalization from the general field of anonymization, clustering from mobility data anonymization, and differential privacy. The approaches are compared on a variety of data sets: New-York city taxis available in open-data, and the Senegal and Cote d'Ivoire data sets available in the scope of the Data for Development (D4D) challenge (de Montjoye et al., 2014; Blondel et al., 2013). Several variations of the data are used to analyze how the various approaches adapt to different conditions, for a total of six distinct data sets.

- As the approaches vary vastly in the format of their output and their interpretations, it is not trivial to compare them fairly. We define and discuss the relevance of various

indicators used to measure the information loss due to anonymization.

The remainder of this chapter is organised as follows: we review related work on data anonymization in Section 3.2. We then detail our methodology in Section 3.3. In Section 3.4, we describe the data sets on which we perform the experiments, and the methods that we compare are listed in Section 3.5. Then we discuss the appropriate indicators to compare the methods in Section 3.6 before presenting our results in Section 3.7. Section 3.8 concludes the chapter.

3.2 State of the art of anonymization

In this section, we first review the k-anonymization methods developed specifically for mobility data, then in the general case of structured data, and finally we review differential privacy, which is another popular principle to anonymize data. We conclude the section with our positioning compared to the k-anonymization techniques from the state of the art.

3.2.1 Trajectory anonymization

Previous work on the anonymization of mobility data has focused on the anonymization of collections of trajectories. The generalization and suppression of complete trajectories is computationally expensive, as it relies on hierarchical clustering with custom merging procedures to accommodate the complex nature of the data. The first notable work is Never Walk Alone (NWA) (Abul et al., 2008), which considers inherent spatial uncertainty to create groups of k individuals that share parts of their uncertainty areas. Its time-tolerant variation, Wait For Me (WMA) (Abul et al., 2010), uses a variation of the edit distance adapted for quantitative values (Chen et al., 2005) in order to handle the uncertainty in time as well as in space. As the edit distance is computationally expensive, its authors also propose a linear spatiotemporal distance intended for large databases. Glove (Gramaglia and Fiore, 2015) uses another rule of merging trajectories that yields better precision and does not add artificial trajectory points in order to create groups of k. In Glove, each point of a generalized trajectory appears as a box containing at least one point of each of the trajectory of the group. When merging two trajectories, the approach associates each generalized point with its closest neighbor to create a larger box (Fig. 3.3). A subsequent approach by Tu et al. (2017) uses another merging logic that also implements l-diversity and t-closeness as well as k-anonymity. Note that a typical value for k in the context of trajectory anonymization would be 2 or 3, with a value of 5 essentially destroying all utility (Gramaglia and Fiore, 2015).

Anonymizing trajectories requires special care on the temporal dimension, which is as important as the spatial ones. For a better interpretability of the output, the concept of **time coherence** is relevant: it states that the time intervals spanned by the generalized points must not overlap. It is possible to interpret an OD matrix as a collection of 2-points trajectories that all start and end at the same time, in which case we can apply these approaches. The merging procedures such as the one used by Glove then become straightforward because of the time coherence constraint: The only possible merge in a single, fixed-timestep OD matrix consists in merging the origins together and the destinations together. These simplifications make the anonymization problem related to clustering, which is a well-studied problem. Yet in this particular problem the expected number of clusters is very large and not fixed in advance, a task for which regular approaches scale poorly (Monath et al., 2021).

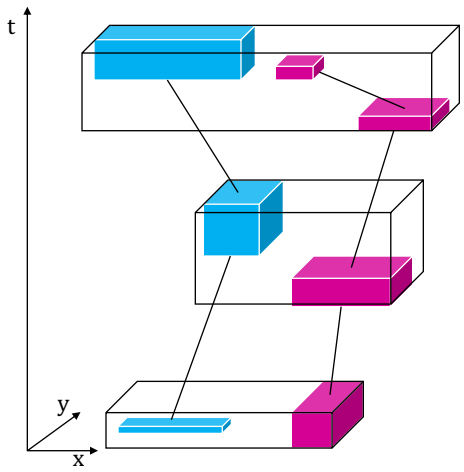


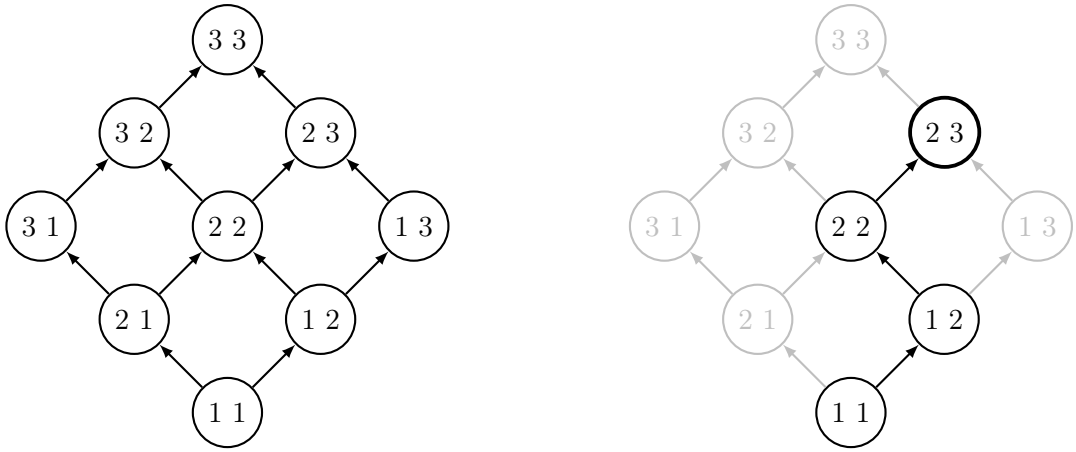
Figure 3.3: Merging two generalized trajectories (one blue on the left, one magenta on the right) in Glove. Each trajectory is defined by its generalized points, which appear as boxes containing the actual spatio-temporal points. The resulting generalized trajectory is given by the three transparent boxes.

3.2.2 Generalization and suppression for relational data

One single OD matrix, corresponding to a single timestep, can be seen as a structured table where each row represents a flow described with only two attributes: the origin o and the destination d. Historically, the first k-anonymization algorithm for structured data is **Datafly** (Sweeney, 2002a), relying on a tree-like Value generalization Hierarchy (VGH) for each attribute (Fig. 3.2). Datafly iteratively generalizes all the values of an attribute to their parent values in the hierarchy. As an heuristic, the particular attribute that is generalized at each step is the one with the most distinct values in the data. This results in a horizontal cut in the hierarchy of each attribute, in the sense that for any given attribute, all individuals are generalized to the same level. We call this family of generalization **uniform**. The uniform generalization approach is still the object of active research: in a graph where each node represents the tuple of the levels of generalizations of the attributes, all possible uniform generalizations form a lattice where we can go from one node to another if they differ by only one for only one of the attributes (Fig. 3.4). Uniform generalization approaches consist in building a path in this lattice, which can be huge and requires an efficient exploration algorithm. Among them, the latest and best adapted to an OD matrix is the **OIGH** algorithm (Mahanan et al., 2021), specifically designed for data where the hierarchy is the same for all attributes.

An example of uniform generalization applied to an OD matrix is given in Fig. 3.5: the initial zones A, B, C, D are associated to a generalization hierarchy in dashed lines, and uniform generalization gives a cut in the tree that aggregates every modality of an attribute to the same level. It is not possible to detail destinations between A and B while aggregating destinations C and D together.

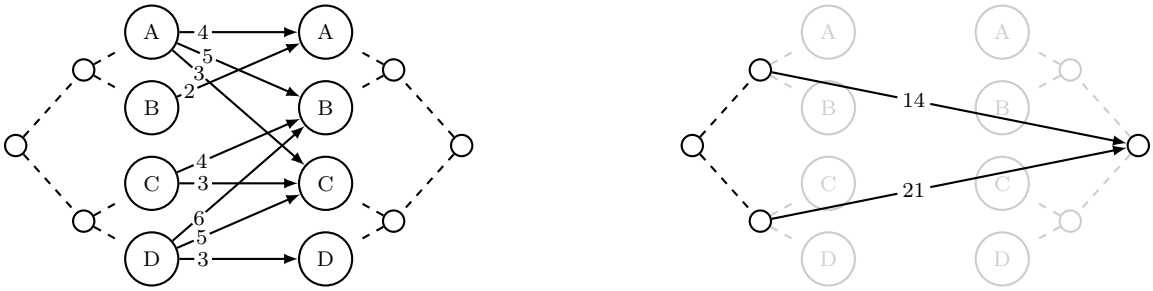
Disregarding the value generalization hierarchy, we can also consider an OD matrix to be composed of four attributes, namely the coordinates of origin and destination. The k-anonymization of quantitative data can be efficiently handled by the **Mondrian** algorithm (LeFevre et al., 2006), which finds a partitioning of attributes using an approach inspired



(a) Lattice of possible generalizations

(b) One path leading to a generalization

Figure 3.4: Left: generalization lattice of a OD-matrix where the maximum level is 3. The first number represents the level of generalization of origins, the second represent the level of generalization of destinations. The original OD-matrix, without generalization, is represented by the node (1 1). Right: Successive generalization steps drawing a path in the lattice, where each direction can be either chosen by an heuristic as in Datafly, or by an exact algorithm as in OIGH.



(a) Initial OD matrix

(b) 10-anonymous OD matrix

Figure 3.5: Example of an OD matrix with hierarchies for origins and destinations, and one possible output of uniform generalization.

by kd-trees (Friedman et al., 1977). Mondrian considers the flows as 4-dimensional points, and at each step selects a dimension and performs a median cut. Then it repeats for each subset of points, and each branch formed this way stops once it is found that any additional cut would imply clusters of less than k points. As such, Mondrian does not allow suppression at all.

In recent works, generalization and suppression have also been formulated as an optimization problem. The constraints of the problem ensure k-anonymity while the objective function to minimize is a measure of the coarseness of generalization. Liang and Samavi (2020) formulate the generalization of each individual value by a mixed-integer linear program. Directly solving for this linear program remains hard, so they propose a practical *Split & Carry* algorithm that splits the problem into smaller, more manageable sub-problems, as well as a greedy search for bigger data sets. The solution found does not necessarily give a partitioning of each attribute as Mondrian does, but rather a partition of the individuals. As such, we can see

their approach as a form of clustering. In the strictest form of k-anonymity, we may accept that an individual is indistinguishable from another if the generalization of their features is included in the generalization of the other's features, without necessarily being equal. This added degree of freedom leads to **non-homogeneous generalization** (Wong et al., 2010), and an application to an OD matrix is illustrated in Fig. 3.6. For more clarity, the data of the figure is summarised in Table 3.1.



(a) Initial OD matrix

(b) Non-homogeneous generalization

Figure 3.6: Left: initial OD matrix to anonymize. Right: same flows with their generalized areas of origin and destination.

volume	initial origin	initial destination	generalized origin	generalized destination
5	C	B	{C, D}	B
4	C	C	C	C
3	D	B	D	B
6	D	C	{C, D}	{B, C, D}
5	D	D	D	{C, D}

Table 3.1: Flows of Fig. 3.6 and their non-homogeneous generalization. Note how any particular initial flow has its origin and destination included in enough generalized flows so that the cumulated volume is above $k = 10$.

Doka et al. (2015) formulate non-homogeneous generalization as a network-flow problem and propose to solve it with an exact algorithm, as well as with a greedy approach that runs in $O(kN^2)$, N being the number of distinct individuals. The Doka approach suffers from this impractical complexity, while proposing an information loss only marginally better than the one obtained by Liang and Samavi (2020). To the author's knowledge, no work has yet formulated generalization and suppression restricted to a generalization hierarchy as an optimization problem.

3.2.3 Differential privacy

Differential privacy is a robust privacy principle introduced by Dwork et al. (2006) that does not apply to a data set in itself but rather to a randomized algorithm that takes the data set as input and returns a set of query results. In our case, the distinction is non-existent as the input data set would be a personal OD matrix and the queries would be the count of all OD flows, which forms an OD matrix.

Principle of differential privacy : The idea of differential privacy is that no individual should have a decisive influence on the results, so that everyone can answer truthfully, knowing that their information won't stand out among the other respondents. This would justify that the result contains insights about a population, but no personal information, and is therefore safe to publish.

More formally, we say that the algorithm respects differential privacy if, for any possible output, the probabilities of returning that output given that a particular individual is or is not in the input data set cannot differ by more than a fixed threshold. The threshold is characterized by a parameter $\epsilon > 0$ called **privacy budget**, which allows to exactly quantify how much privacy we want to enforce. This definition is formalized in Eq. 3.1, stating that a randomized algorithm $\mathcal{M}$ is ϵ-differentially private if for all subset $\mathcal{S}$ of the possible values returned by $\mathcal{M}$, and for all couples (x, y) of data sets differing by at most one row:

$$\Pr[\mathcal{M}(x) \in \mathcal{S}] \leq \exp(\epsilon)\Pr[\mathcal{M}(y) \in \mathcal{S}]. \tag{3.1}$$

Where $\Pr[\mathcal{M}(x) \in \mathcal{S}]$ and $\Pr[\mathcal{M}(y) \in \mathcal{S}]$ are the probabilities of $\mathcal{M}$ returning a value belonging to $\mathcal{S}$ given the input data sets x and y, respectively. Note that the factor $\exp(\epsilon)$ is greater than 1 for any $\epsilon > 0$, giving us a two-sided bounding of the ratio of probabilities.

Differential privacy can be seen as a formalization of the principle of uninformativeness: no individual can alone have a decisive influence on the outcome, so we cannot learn the individual information from the results.

Differential privacy applied to mobility data : As differential privacy is only a property, there are several ways to implement algorithms with this property, and they may differ with respect to the nature of the data set and queries at hand.

A first remark is that differential privacy is applicable to trajectory data: In that case, the noise is applied either to the trajectory points themselves (Jiang et al., 2013), or to other metrics of the trajectories (Jin et al., 2022), followed by a modification of the data in order to match the new target metrics.

In our case, an OD matrix falls into the category of **histogram queries**, where we query the original data for the count of every OD flow. The most straightforward way of satisfying differential privacy for histogram queries is to add a Laplacian noise to each count (Dwork and Roth, 2014) (Fig. 3.7).

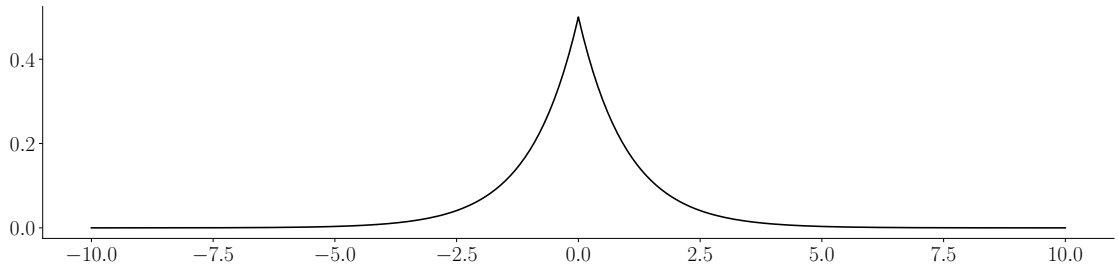


Figure 3.7: Distribution of the Laplace law of mean 0 and scale 1.

The amplitude of the noise depends on the desired privacy budget ϵ, and on the **sensitivity**

of the queries. In this context, we call sensitivity of a query the amount by which the result of a query can change when one row of the data set changes. The sensitivity of histogram queries is one, because an individual can only be in one bin of the histogram and therefore only count for one in the sum of the bins. For a default privacy budget $\epsilon = 1$, differential privacy is thus obtained by adding a Laplacian noise of parameter $\frac{1}{\epsilon} = 1$ to each OD flow. As it would not make sense to have negative or non-integer flows, we round the flows and keep only the positive ones, without hampering the differential privacy guarantee (Cormode et al., 2011). In fact, differential privacy has the convenient property of being immune to post-processing (Dwork and Roth, 2014), meaning that we can perform any treatment we want on the output of an ϵ-differentially private algorithm, the result will never be less than ϵ-differentially private. This property is consistent with the GDPR recommendation that anonymization should be irreversible. The behavior of the resulting noise process is illustrated in Fig. 3.8, which shows the distribution probability of the output values for 4 different input volumes.

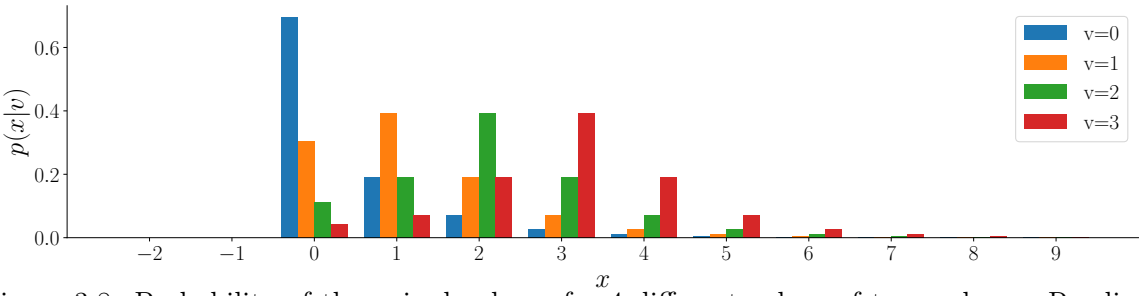


Figure 3.8: Probability of the noised volume for 4 different values of true volume. Reading: the probability that a flow of volume 1 is shown as a flow of volume 0 after the noise process is around 0.3, while it is only around 0.1 if the true volume is 2.

It is important to note that in order to satisfy the differential privacy criterion even for users who would be alone in their flow, we must also apply noise to 0-flows. The probability that our noise process will return a non-zero noisy flow given a true flow of 0 is 30.33%. OD matrices are known to be very sparse, and the density of our data sets ranges from 0.24% to 4.4%. As a result, a differentially private OD matrix would be between 6 and 126 times more expensive to store than the original matrix (Cormode et al., 2011), and would consist mostly of noise applied to a constant signal.

Differential privacy and GDPR : There is a fundamental disconnect between the philosophy of differential privacy and the GDPR criteria: Differential privacy starts with an abstract definition of what privacy should be, and concludes that the addition of a small amount of noise would be sufficient to meet that definition. In that sense, the fact that the noise is small is a desirable property of the anonymization process. In our case, the really small noise has a 40% probability of being 0, and an 80% probability of being between -1 and $+1$. As we also clip negative values to 0, the probability of a small flow being unchanged by our adapted Laplacian mechanism is actually a little more than 40%.

From the GDPR point of view the addition of noise is an acceptable way to anonymize the data, but its relevance would be assessed based of the actual changes it makes in the data set instead of an *a priori* property of the noise. The fact that the Laplacian mechanism leaves 40% of the small flows unchanged is a very negative property of the anonymization process, as it

means that a respondent who happens to outlier will likely be exposed without any protection. If the outlier happens to benefit from a small, positive noise, then the processed data could show several people instead of the one outlier, but this does little to reassure the respondent that their data are safe.

The apparent solution to guarantee a noise level compliant with the GDPR would be to tune the privacy budget ϵ to increase the amplitude of the noise. As doing so risks degrading the quality of the data, we could also look for a modification on the origin and destination regions so that the flows may be larger while keeping a noise of same amplitude. In either case we could not really claim that we follow the principles of differential privacy, which imply fixing a privacy budget and deriving a sufficient noise to guarantee property 3.1. Instead, what we are doing when tuning the privacy budget ϵ is that we have criteria fixed by our definition of anonymization, and we are looking for a noise that would happen to meet these criteria. The fact that this noise happens to guarantee property 3.1 is purely incidental.

In a more general setting, the fact that differential privacy is only an individual property is a strong limitation on the guarantees it can provide for the resulting data sets. A pessimistic view of the differential privacy property given in Eq. 3.1 would be that it only ensures that the study will not be significantly more harmful to a user's privacy if they participate than if they don't. It provides no guarantee against probabilistic attacks, since they do not require the target to participate in the study in order to be successful. In fact, allowing probabilistic attacks is an explicit feature of differential privacy (Dwork and Roth, 2014). The argument is that, with medical data in mind, the quality of a data set is measured by the correlations we can find between the attributes. The most common example given is a data set that allows us to find that people who smoke have a higher risk of lung cancer. In the differential privacy philosophy, such a data set is not problematic as long as the conclusion does not depend on any single person participating in the study. However, if we take as an example a survey asking where pregnant Texan women cross the border into New Mexico to get an abortion, then even if the outcome of the survey is independent of any individual participant, it is clear that one participant could potentially endanger others, and that the resulting data set is not safe to publish. Note that k-anonymity alone would also not be a satisfactory property in this case. It merely illustrates that the conditions under which a data set is safe to publish should refer to the data set in itself, and not to an individual-level property.

In conclusion, differential privacy provides a rigorous framework for dealing with private data, with a focus on the utility we could derive from a medical data set where correlations are important. In the case of very sparse data, such as OD matrices, the application of differential privacy is technically questionable and conceptually at odds with GDPR, whose constraints we explicitly seek to satisfy in this study. In fact, investigations into its relevance with respect to the GDPR (Holzel, 2019) find that differential privacy does not guarantee the safety of the data set in itself, nor does it protect against false re-identifications, which should be prevented as well as true ones. For more practical details on differential privacy, see (Team, 2020).

3.2.4 Positioning of our anonymization method

The presented solutions tend to focus on finding the finest generalization that respects k-anonymity, while keeping a reasonable computing time for small-scale data sets. Yet mobile

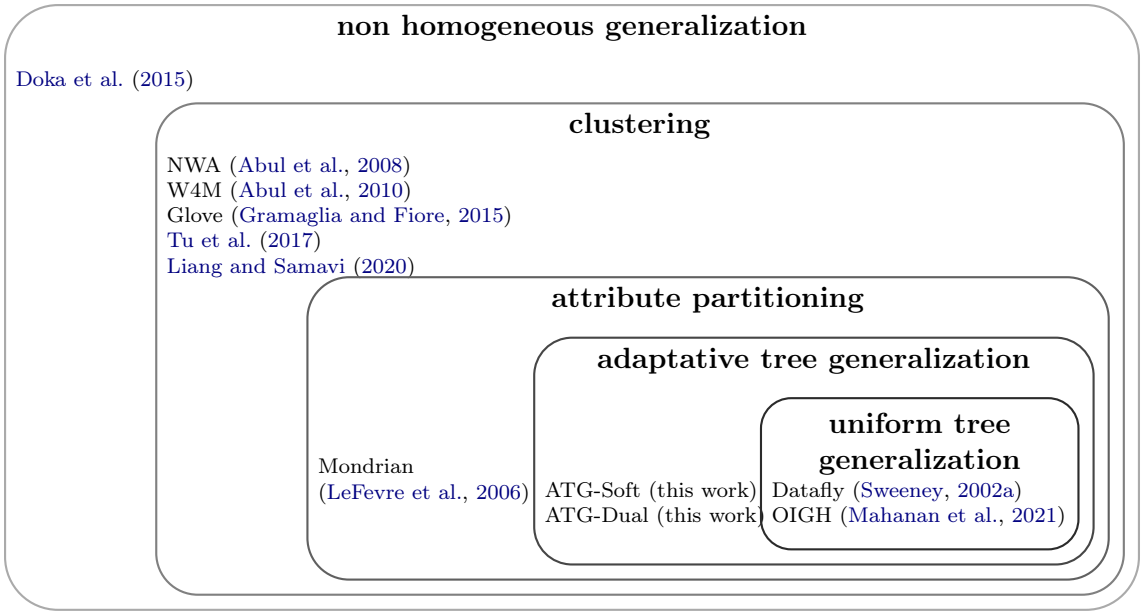


Figure 3.9: Classification of k-anonymization approaches with respect to the space of solutions they consider.

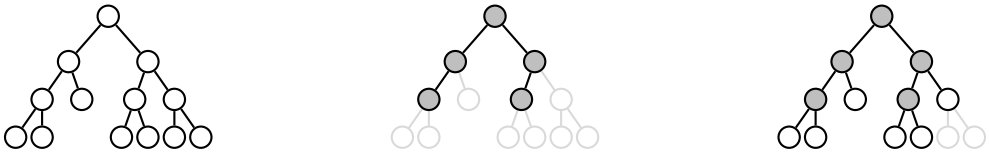
phone data offer volumes way above the ones usually considered, to which few solutions from the state of the art can actually scale. For example, for a data set of 100,000 records, 4 attributes and $k = 5$, *Split & Carry*, which is the fastest model proposed by Liang and Samavi (2020), takes hours to complete: the computing time being exponential with respect to k, it may amount to days for $k = 10$. These time scales are orders of magnitude above the computing time we aim for in the benchmark. We propose in this paper a simplified context for the optimization problem, which performs an Adaptative Tree generalization specifically adapted for OD matrices, with their particularity of being exceptionally low-dimensional, high-volume, and featuring a high number of modalities. Our **ATG-Dual** approach can handle high volumes while being significantly finer than the faster approaches of uniform generalization. We also propose **ATG-Soft**, an even lighter version that sacrifices precision in favor of computing time. Fig. 3.9 summarises the positions of the approaches of the state of the art with respect to the degree of freedom they allow in the solution: uniform tree generalization restricts the solution space the most, followed by our approach and then Mondrian, which all find partitions of the attributes in order to form clusters. Methods that find a clustering that does not necessarily rely on a partitioning of the attributes have a broader search space. Finally, non-homogeneous generalization, which does not necessarily produce a partitioning of the individuals, has an even broader search space.

3.3 Anonymization methodology

3.3.1 Overview

We consider an OD matrix, defined on a set of initial tiles and equipped with a generalization hierarchy that describes a local order in which we can aggregate the initial tiles to form generalized areas. We call the hierarchy T, and we use the same hierarchy over the origins and the destinations. Each leaf of T represents an initial tile, and each node $n \in T$ represents the

area obtained by the union of the children of n. A pruning of T is a tree obtained by pruning out some branches of T (*i.e.,* removing a node and all the nodes below it). We interpret such a pruning as the set of nodes we wish to **split**, so that their children, if not split, represent a partitioning of the set of initial tiles. This is illustrated in Fig. 3.10: the pruning defined by the nodes in grey is interpreted as a partitioning defined by the nodes in white in the right-most figure.



(a) Complete hierarchy (b) Nodes selected by pruning (c) Interpretation as a partitioning

Figure 3.10: Left: Example of a complete generalization hierarchy. Middle: Example of a pruning of the hierarchy, satisfying the tree constraint. Right: Interpretation of the pruning as a partitioning of the values.

Formally, we note $(x_n)_{n\in T}$ the binary variable that describes a pruning of T. In this section, we will use the **tree constraint**, that holds if and only if the set of nodes n such that $x_n = 1$ is a pruning of T:

$$\text{tree constraint}(\mathbf{x}, T) = \begin{cases} \forall n \in T, x_n \in \{0, 1\} \\ \forall n \in T, x_{p(n)} \geq x_n \\ \forall n \in leaves(T), x_n = 0 \\ x_{p(root(T))} = 1 \text{ by convention} \end{cases}$$

where $p(n)$ denotes the parent of node n in the tree T. As we cannot split leaves of the hierarchy, x_n is necessarily 0 for leaves of T. Let $\mathcal{O}$ be a partitioning of the initial origin tiles, and for each $o \in \mathcal{O}$, $\mathcal{D}_o$ be a partitioning of destination tiles. For each $o \in \mathcal{O}$, $d \in \mathcal{D}_o$, we note $|o|$ and $|d|$ the sizes of the origin and destination, measured in number of initial tiles and $o \to d$ the flow going from o to d. We note $v_{o\to d}$ the volume of the flow $o \to d$, measured in number of people and obtained by summing all the flows from one initial origin belonging to the aggregated origin o to one initial destination belonging to the aggregated destination d. We consider the **total generalization error** G, derived from the objective functions used in the state of the art (Doka et al., 2015; Liang and Samavi, 2020):

$$G = \sum_{o\in\mathcal{O}} \sum_{\substack{d\in\mathcal{D}_o \\ v_{o\to d}\geq k}} (|o| + |d|)v_{o\to d}. \tag{G}$$

The generalization error G penalises the size of the origin and destination for each individual that has not been suppressed. As we have to suppress any individual that has not been k-anonymized, the amount of suppression is given by:

$$S = \sum_{o\in\mathcal{O}} \sum_{\substack{d\in\mathcal{D}_o \\ v_{o\to d}<k}} v_{o\to d}. \tag{S}$$

We now consider the problem of finding a partitioning of origins and of destinations that minimizes G while keeping S under a suppression constraint C. The formulation of this problem is given for reference:

Problem (Coupled generalization under maximal suppression constraints)**.** Let $\mathbf{x} = (x_o)_{o\in T}$ represent an origin's pruning, and for each possible origin $o \in T$, $\mathbf{y_o} = (y_{o,d})_{d\in T}$ represent one destination's pruning. Then the coupled partitioning problem under maximal suppression constraint can be cast as:

$$\min_{x,y} \sum_{\substack{o\in T, d\in T \\ v_{o\to d}\geq k}} (x_{p(o)} - x_o)(y_{o,p(d)} - y_{o,d})(|o| + |d|)v_{o\to d}$$
$$s.t. \begin{cases} \text{tree constraint}(\mathbf{x}, T) \\ \forall o, \text{ tree constraint}(\mathbf{y_o}, T) \\ \sum\limits_{\substack{o\in T, d\in T \\ v_{o\to d}< k}} (x_{p(d)} - x_d)(y_{o,p(d)} - y_{o,d})v_{o\to d} \leq C \end{cases} \tag{J}$$

Note that the term $(x_{p(o)} - x_o)$ ensures that we only count the penalties of the leaves of the origin pruning, and the term $(y_{o,p(d)} - y_{o,d})$ ensures that we only count the penalties of the leaves of the destination pruning.

Problem J is already a restriction on the problem of generalization of flows, yet it remains hard and computationally expensive to solve. In order to simplify it, we uncouple it in two separate problems: we first formulate the **best pruning problem** to find a map of origins $\mathcal{O}$ with the heuristic that origins $o \in \mathcal{O}$ should emit an outgoing volume close to a given target volume v_{target}. As the k-anonymization is coarser for problems with fewer individuals, this heuristic ensures that the destination maps for different origins will be close in coarseness. It is best justified with the assumption that the flows have similar behaviors regardless of their origins. Then, for each origin $o \in \mathcal{O}$, we can more easily find a partitioning of destinations that minimizes the total generalization error G under the suppression constraint. We detail Algo. 1, to decouple the generalization problem. It relies on generalizing the origins by solving Problem P with function `get_pruning` (detailed further in Algo. 2), and then generalizing the destinations by solving a similar problem with an additional constraint on the volume we are allowed to suppress. These problems are formulated as variations of the best pruning problem, which is described in the next section.

3.3.2 Best pruning problem

Formulation: The best pruning problem considers that generalizing values to a node n induces a penalty π_n. Finding the best partitioning then amounts to minimizing the sum of penalties of the leaves of the pruned tree. The resulting problem is given by:

Algorithm 1: Decoupling of the generalization problem

Input : $v_{o\to d}$, hierarchy_tree, v_target, k, S
Output: aggregated_od_matrix

```
od_matrix_agg ← ∅
error_fun ← (f : o ↦ (∑_d v_{o→d} − v_target)^2)
_, generalized_origins ← get_pruning(hierarchy_tree, error_fun)
agg_flows ← generalize_destinations(generalized_origins, hierarchy_tree, k, S)
for (o, d, vol) ∈ agg_flows do
    od_matrix_agg.append((o, d, vol))
end for
return od_matrix_agg
```

Problem (Best pruning problem)**.** Let x represent a pruning of T, then the best pruning problem for the penalty function π is defined by:

$$\begin{aligned} &\min_{x} \sum_{n\in T} (x_{p(n)} - x_n)\pi_n \\ &s.t.\ \text{tree constraint}(\mathbf{x}, T) \end{aligned} \tag{P}$$

The term $(x_{p(n)} - x_n)$ ensures that we only count the penalties of the leaves of the pruned tree.

The most important feature of π_n is that the value of the penalty associated to a node n does not depend on whether other nodes have been split or not, and in particular does not depend on nodes outsides of the branch starting at n. This propery makes solving problem P rather straightforward: by making a simple pass through T, one can recursively choose for each node if it is better to split it or not, with the corresponding best-case penalty. It is possible that a node n has a minimum penalty by being split (*i.e.,* $x_n = 1$), but that its parent does not. In that case, as x is required to satisfy the tree constraint, the node n is not split.

Solving the best pruning problem: We detail Algo. 2 to solve the Best Pruning Problem P: For each node n, α_n can be evaluated in time $\mathcal{O}(1)$. If we note M the number of leaves of T, which correspond in our case to the size of the study area, then a tree has $\mathcal{O}(M)$ nodes and solving problem P can be done in time $\mathcal{O}(M)$. Problem P can also be formulated as a maximisation of the benefit compared to splitting nothing, in which case the objective function sums over all selected nodes instead of only the leaves. This other form of the problem is known as the maximal closure problem, and can also be solved using an efficient max-flow/min-cut algorithm (Picard, 1976).

Generalization of the origins with the best pruning problem: In the case of generalization of origins, we set π_n so that it measures how far the volume of flows coming out of n is from our target volume v_{target}. We choose the expression:

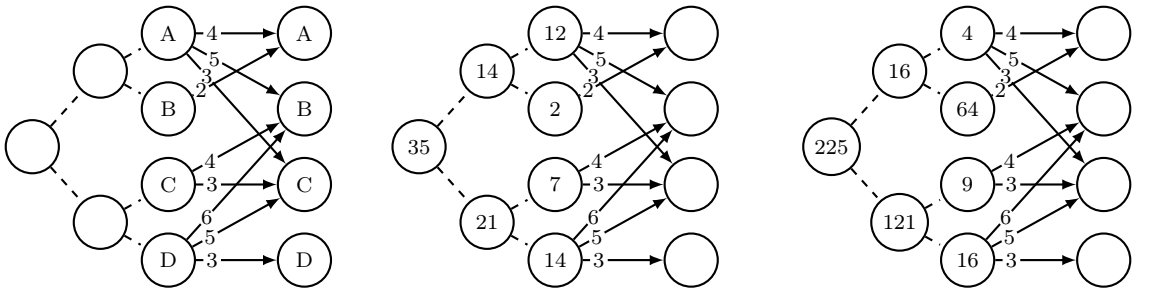
$$\pi_n = \left(v_{target} - \sum_{d\in leaves(T)} v_{n\to d} \right)^2 .$$

Algorithm 2: get_pruning

Input : node n, error_fun
Output: best pruning error of node n, and the corresponding pruning

```
error_agg = error_fun(n)
if tree has children then
    error_split ← 0
    pruning_split ← ∅
    for child ∈ tree.children do
        bpe_c, pruning_c ← get_pruning(child, error_fun)
        error_split ← error_split + bpe_c
        pruning_split ← pruning_split ∪ pruning_c
    end for
    if error_split < error_agg then
        return error_split, pruning_split
    else
        return error_agg, {n}
    end if
else
    return error_agg, {n}
end if
```

An illustration of the corresponding values for π_n is given in Fig. 3.11. In this illustration we see that for each node, we can choose either to accept its penalty or accept the sum of the penalties of its children.



(a) Initial OD matrix (b) Emission flows (c) Leaf penalties ($v_{target} = 10$) $\pi_n = (v_{target} - v_n)^2$

Figure 3.11: Left: example of an OD matrix with the hierarchy over the origins. Middle: emission flows from each initial and generalized origin. Right: corresponding penalty obtained by setting $\pi_n = \left(v_{target} - \sum_{d \in leaves(T)} v_{n \to d}\right)^2$ with $v_{target} = 10$.

Then, we use a variation of Problem P to find a generalization of destinations that minimizes the total generalization error G given a fixed $\mathcal{O}$, under the suppression constraint C.

3.3.3 Pruning problem with hard suppression constraint

In this section, we formalize the best pruning problem under the constraint that failing to generalize some destinations can lead to flows under the volume k that then have to be suppressed. The total volume of the suppressed flows is required to be under the constraint

threshold C, which introduces a variation on the best pruning problem described above.

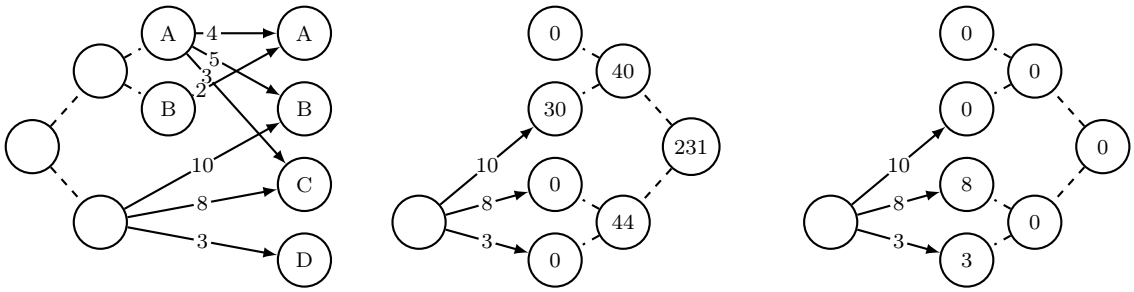
Formulation: Let $o \in \mathcal{O}$ be an origin. For each $d \in T$, we define the **aggregation penalty** α_d, equal to the contribution of node d in the expression of G:

$$\alpha_d = \begin{cases} (|o| + |d|)v_{o\to d} & \text{if } v_{o\to d} \geq k \\ 0 & \text{else} \end{cases}, \tag{3.1}$$

and the **suppression penalty** σ_d, equal to the number of records that would be suppressed if d was a leaf of the pruned tree:

$$\sigma_d = \begin{cases} 0 & \text{if } v_{o\to d} \geq k \\ v_{o\to d} & \text{else} \end{cases}. \tag{3.2}$$

Examples of values for α_d and σ_d are given in Fig. 3.12: the size of the origin is 2 as it is the result of aggregating the two areas C and D together. We see that the values of α_d increase as we go up in the hierarchy. The lowest total aggregation penalty is achieved by aggregating nothing, but it may imply suppressing more volume than is acceptable.



(a) Initial OD matrix (b) Aggregation penalties (c) Suppressed volumes ($k = 10$)

Figure 3.12: Left: example of an OD matrix with the hierarchy over the destinations. Middle: aggregation penalty for each destination for one given origin. Right: suppressed volumes for each destination, considering $k = 10$.

Minimizing G with fixed origin o amounts to minimizing the sum of α_d while keeping the sum of σ_d under a suppression constraint C. The resulting problem is given by:

Problem (Best pruning problem under hard suppression constraints)**.** Let x represent a pruning of T and α_d and σ_d be defined according to Equations 3.1 and 3.2 respectively, then finding the best pruning of T for a fixed origin o is equivalent to:

$$\min_x \sum_{d\in T} (x_{p(d)} - x_d)\alpha_d$$
$$s.t. \begin{cases} \text{tree constraint}(\mathbf{x}, T) \\ \sum_{d\in T} (x_{p(d)} - x_d)\sigma_d \leq C \end{cases} \tag{H}$$

Like in Problem P, the term $x_{p(d)} - x_d$ ensures we only count the aggregation and sup-

pression penalties of the leaves of the pruned tree. Then, splitting a node d brings a direct gain b_d in terms of aggregation penalty, as the sum of the α_c for all children c of the node d cannot be greater than α_d. Likewise, splitting d can incure a cost w_d as it may be necessary to suppress some flows going to the children of d. As such, each node can be associated to a gain and a cost, which can be used to reformulate Problem H as a knapsack problem, with an additional constraint due to the tree structure. Reformulating Problem H as a knapsack problem is a formalization of the idea that, in our top-down generalization approach, we want to select which nodes to split.

Problem (Knapsack formulation of the best pruning problem under hard suppression constraints)**.** For each node d, let b_d and w_d be defined as:

- $b_d = \alpha_d - \sum_{c \in children(d)} \alpha_c$ be the gain due to splitting d,
- $w_d = \sum_{c \in children(d)} \sigma_c - \sigma_d$ the weight of destination d.

Then the knapsack formulation of H is:

$$\begin{aligned} &\max_x \sum_{d \in T} x_d b_d \\ &s.t. \begin{cases} \text{tree constraint}(\mathbf{x}, T) \\ \sum_{x \in T} x_d w_d \leq C \end{cases} \end{aligned} \tag{K}$$

With this formulation, the problem has been studied under the name of the Tree knapsack problem (van der Merwe and Jacobus, 2007; Shaw and Cho, 1998) or ordered knapsack problem (Johnson and Niemi, 1983). A variant where the hierarchy is not necessarily a tree is known as the **Precedence Constraint Knapsack Problem (PCKP)** (Kellerer et al., 2004), and has been extensively studied in the field of open-pit mining (Byun and Dimitrakopoulos, 2013; Maiti et al., 2021). The tree knapsack problem can be easily solved by dynamic programming for trees under a few hundred nodes, but for larger trees it is recommended to consider its dual (Byun and Dimitrakopoulos, 2013). Equivalently, we directly consider the dual of our original formulation H.

Problem (Dual formulation of best pruning problem under hard suppression constraints)**.** The dual of problem H, obtained by relaxation of the suppression constraint is given by:

$$\begin{aligned} &\max_\lambda \min_x \sum_{d \in T} (x_{p(d)} - x_d)(\alpha_d + \lambda \sigma_d) - \lambda C \\ &s.t. \text{ tree constraint}(x, T) \end{aligned} \tag{D}$$

Solving the best pruning problem under hard constraint: Problem D is a variation on the dual of the PCKP, as we obtain a maximisation problem instead of a minimization problem. We can still solve it with the **All Breakpoints Algorithm** (Byun and Dimitrakopoulos, 2013): This algorithm relies on the fact that the lagrangian dual function that we aim to

maximize:

$$g : \lambda \mapsto \min_x \sum_{d \in T} (x_{p(d)} - x_d)(\alpha_d + \lambda \sigma_d) - \lambda C, \tag{3.2}$$

is continuous and piecewise linear, as illustrated in Fig. 3.13 (left).

$G(\lambda)$ is piecewise linear: We can prove that the best pruning error $g(\lambda)$ is piecewise linear: As the $(x_n)_{n \in T}$ are in $\{0, 1\}$, the solution x for $G(\lambda)$ is necessarily constant in a neighborhood of λ. Then, we see from Eq. 3.2 that for a fixed value of $(x_n)_{n \in T}$, $g(\lambda)$ is linear with respect to λ. This linear relation is only broken if the solution $(x_n)_{n \in T}$, change values, which can only happen for a finite number of values for λ and forms a new piece of the linear function. These transition points are called **breakpoints** and correspond to two distinct solutions that give the same value $g(\lambda)$.

Computing $g(\lambda)$ via the best pruning error: This list of breakpoints can be recursively computed by a single pass through T: it suffices for each node d to define the problem D on the branch starting from d, and find the breakpoints of its lagrangian dual function. We call this auxiliary lagrangian dual function the **best pruning error $B_d(\lambda)$** of d. The best pruning error of the root $B_{root(T)}$ is then the lagrangian dual function g. We introduce the **best pruning error $B_d(\lambda)$** of d, which is equal to the term d of the sum in the expression of $g\lambda$ (Eq. 3.2). By construction, as a node d has the choice of either be split or not, with a different penalty in each case, $B_d(\lambda)$ is equal to the minimum penalty between splitting the node and not splitting it:

$$B_d(\lambda) = \begin{cases} \min \left\{ \sum_{c \in \text{children } (d)} B_c(\lambda), \alpha_n + \lambda \sigma_d \right\} & \text{if } d \text{ has children and } v_d \geq k \\ \alpha_d + \lambda \sigma_d & \text{else} \end{cases} \tag{3.3}$$

The best pruning error of a leaf l can be computed in constant time: it is either constant equal to α_l or linear with slope σ_l. It has no breakpoints. The best pruning error $B_d(\lambda)$ of any node d can be computed from the best pruning errors of its children using Eq. 3.3. The sum of the best pruning error means that, unless some breakpoints of the children functions coincide, the best pruning error of d has a number of breakpoints equal to the sum of the number of breakpoints of each child. The minimum condition acts as a ceiling threshold at large values of λ, and implies that there is at least one more breakpoint in $B_d(\lambda)$ than in the sum of the best pruning errors of the children. By solving a simple inequality to find the value of λ associated to this additional breakpoint, and storing the parameters of all the linear pieces, we can recursively compute the best pruning error of the root of T. By induction, $B_{root(T)}(\lambda)$ is the minimal value for the generalization penalty $\alpha_d + \lambda \sigma_d$, and that the Lagrangian function for the original Problem D is given by:

$$g(\lambda) = B_{root(T)}(\lambda) - \lambda C.$$

$B_d(\lambda)$ is non-decreasing: As increasing λ means adding more penalization, the best solution x can only have a higher value. It follows that B_d is non-decreasing with respect to λ. It can be constant, and for each d such that $v_d \geq k$ it actually assumes a constant value after a threshold λ_d^*. Indeed, intuitively, if the penalisation for suppression λ is high enough, it is

preferable not to split node d, and accept the constant aggregation penalty α_d. If node d has a volume $v_d < k$, then it can only yield the suppression penalty $\lambda\sigma_d$ and its best pruning error is a linear function.

$B_d(\lambda)$ is concave: Finally, We prove by induction that for each $d \in T$, the best pruning error B_d is concave with respect to λ:

- If d is a leaf of T, then from the expression of B_d it is either constant or linear, so it is concave.
- For a given $d \in T$, B_d is the sum of the best pruning error of the children, which are supposed concave. Then it is maxed by a threshold, which conserves concavity.

It follows that $G : \lambda \mapsto B_{root(T)}(\lambda) - \lambda C$ is also concave.

The solution of the dual still respects the suppression constraint We can prove that the optimal solution x^* of problem D is feasible for problem H, meaning that the solution respects the hard constraint of no more than C records suppressed:

$$\sum_{x \in T}(x_{p(d)} - x_d)\sigma_d \leq C.$$

As illustrated above, the function $g : \lambda \mapsto \min_x \sum_{d \in T}(x_{p(d)} - x_d)(\alpha_d + \lambda\sigma_d) - \lambda C$ is piecewise linear, characterized by its breakpoints. Each piece of g is defined on bounds $[\lambda_l^0; \lambda_r^0]$ and represents at least one solution

$$x^0 = \underset{x \in [\lambda_l^0; \lambda_r^0]}{\operatorname{argmin}} \sum_{d \in T}(x_{p(d)} - x_d)(\alpha_d + \lambda\sigma_d) - \lambda C.$$

The value associated to x^0 depends linearly on λ, with derivative:

$$\forall \lambda \in [\lambda_l^0; \lambda_r^0], L'(\lambda) = \sum_{d \in T}(x^0_{p(d)} - x^0_d)\sigma_d - C.$$

Two consecutive pieces of g share a single point $(\tilde{\lambda}, G(\tilde{\lambda}))$, meaning the solution x^0 corresponding to the left piece and the solution x^1 corresponding to the right piece have the same value $g(\tilde{\lambda})$. As proven before, g is concave. Moreover, $B_{root(T)}$ is a constant value after a threshold $\lambda^*_{root(T)}$, assuming the trivial property $v_{root(T)} \geq k$. It follows that $G(\lambda)$ is necessarily decreasing after this $\lambda^*_{root(T)}$. As it is defined for $\lambda \geq 0$, we conclude that the maximum of G is reached. It is necessarily reached for a least one breakpoint $(\lambda^*, g(\lambda^*))$, which necessarily verifies: $g'_-(\lambda^*) \geq 0$ and $g'_+(\lambda^*) \leq 0$, where G'_- and G'_+ denote the left-hand and the right-hand derivatives, respectively. Then the derivative of the right piece is $g'_+(\lambda^*)$, so the solution x^* corresponding to the right-hand piece of g verifies:

$$\sum_{d \in T}(x^*_{p(d)} - x^*_d)\sigma_d - C \leq 0.$$

We proved that there always exist an optimal solution x^* implied by the dual problem D that respect the suppression constraint.

The Specific Breakpoints Algorithms: We discussed in the above paragraphs that the best pruning error B_d of a given node d can be determined by a single pass through the branch of the tree starting at d. However the computing time also depends on the number of its breakpoints, which can be high. The **Specific Breakpoints Algorithm (SBA)** (Maiti et al., 2021) improves over the All Breakpoints Algorithm: It takes advantage of the piecewise linearity and convexity (in this case, concavity) of the lagrangian dual function to perform a search akin to dichotomy, which requires evaluating the lagrangian dual function $G(\lambda)$ for only a small number of values of λ (Fig. 3.13, right).

Algorithm 3 is an adaptation of the Specific Breakpoints Algorithm for the leaf-error minimization, instead of maximisation. The Specific Breakpoints Algorithm requires to evaluate the value and the derivative of the best pruning error, which is done in a similar way than the solving of the pruning in Algo. 2. As it does not return the same output, we include it in a separate algorithm (Algo. 4).

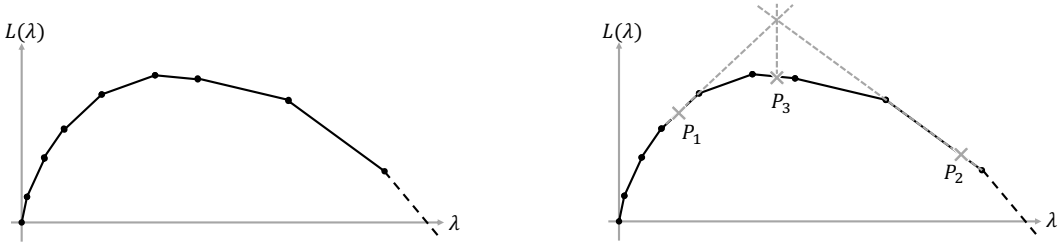


Figure 3.13: Left: General shape of the lagrangian dual function of problem D. Right: Geometrical interpretation of one step of the algorithm SBA used to find the maximum.

3.3.4 Pruning problem with soft suppression constraint

Evaluating the Lagrangian $g(\lambda)$ for a single value λ is equivalent to solving problem P with $\pi_d = \alpha_d + \lambda\sigma_d$, eventually adding $-\lambda C$. We call this variation the soft suppression constraint, which is formulated as:

Problem (Best pruning problem with soft suppression constraint)**.**

$$\begin{aligned} &\min_{x} \sum_{d \in T} (x_{p(d)} - x_d)(\alpha_d + \lambda\sigma_d) - \lambda C \\ &s.t.\ \text{tree constraint}(x, T) \end{aligned} \quad . \tag{H'}$$

Note that this is also closer to minimizing a common definition of the generalization error G (Liang and Samavi, 2020), which contrary to our definition (Eq. G), penalizes suppressed records as if they were generalized to the maximum level. Here, suppressed records are penalized as if they were generalized to the level λ, and we can interpret λ as an area. The optimal solution for problem H' has an interesting property: as we prove in the next paragraph, it will never imply the generalization of a flow $o \to d$ such that $|o| + |d| > \lambda$. As such, we lose the hard constraint on the suppressed volume but gain a hard constraint on the level of generalization, which can also be of interest for data owners.

Algorithm 3: generalize_destinations

Input : dest_tree, k, C
Output: aggregated_od_matrix

$\lambda_l \leftarrow 0$
bpe$_l$, bpe'$_l \leftarrow$ dest_tree.root.evaluate(λ_l)
$e_l \leftarrow$ bpe$_l - \lambda_l C$ `// value of objective function`
$s_l \leftarrow$ bpe'$_l - C$ `// derivative of objective function`
$\lambda_r \leftarrow$ M
bpe$_r$, bpe'$_r \leftarrow$ evaluate(dest_tree.root, λ_r)
$e_r \leftarrow$ bpe$_r - \lambda_r C$
$s_r \leftarrow$ bpe'$_r - C$
while True **do**
  $\lambda_m \leftarrow (s_l\lambda_l - s_r\lambda_r + e_r - e_l)/(s_l - s_r)$ `// find the intersection of the tangents`
  bpe$_m$, bpe'$_m \leftarrow$ evaluate(dest_tree.root, λ_m)
  $e_m \leftarrow$ bpe$_m - \lambda_m S$
  $s_m \leftarrow$ bpe'$_m - C$
  if $\lambda_m = \lambda_l$ **or** $\lambda_m = \lambda_r$ **then**
    error_fun $\leftarrow (f : n \mapsto \alpha_n + \lambda_m \sigma_n)$
    return get_pruning(dest_tree, error_fun)
  else
    if $s_m > 0$ **then**
      $(\lambda_l, e_l, s_l) \leftarrow (\lambda_m, e_m, s_m)$
    else
      $(\lambda_r, e_r, s_r) \leftarrow (\lambda_m, e_m, s_m)$
    end if
  end if
end while

Algorithm 4: evaluate

Input : node n, λ
Output: best pruning error of node n, and its derivative w.r.t λ

```
error_agg = α_n + λσ_n
deriv_agg = σ_n
if tree has children then
    error_split ← 0
    deriv_split ← 0
    for child ∈ tree.children do
        bpe_c, bpe'_c ← evaluate(child, λ)
        error_split ← best_error_split + bpe_c
        deriv_split ← best_deriv_split + bpe'_c
    end for
    if error_split < error_agg then
        return error_split, deriv_split
    else
        return error_agg, deriv_agg
    end if
else
    return error_agg, deriv_agg
end if
```

The soft suppression constraint introduces a hard generalization constraint: In this paragraph, we prove that any volume would rather be split than aggregated to a level above λ, even if it means getting entirely suppressed. This does not concern flows whose destination cannot be split (leaves of T), or flows that are already suppressed ($v_{o\to d} < k$). This can be formalized as:

$$\forall d \in T \sim \mathit{leaves}(T) \text{ s.t. } v_{o\to d} \geq k, |o| + |d| > \lambda \implies x_d^* = 1.$$

Proof. Let $\lambda \geq 0$, $d \in T$, $o \to d$ a flow such that $|o|+|d| > \lambda$. We note $\mathcal{C}$ the set of the children of d. We suppose d is not a leaf of T, so $\mathcal{C} \neq \emptyset$. Recall that in the context of OD aggregation, the aggregation penalty of node d is:

$$\alpha_d = \begin{cases} (|o| + |d|)v_{o\to d} & \text{if } v_{o\to d} \geq k \\ 0 & \text{else} \end{cases},$$

and the suppression error of node d is:

$$\sigma_d = \begin{cases} 0 & \text{if } v_{o\to d} \geq k \\ v_{o\to d} & \text{else} \end{cases},$$

so the best pruning error of node d, defined by:

$$B_d(\lambda) = \begin{cases} \min\left\{ \sum_{c\in\mathcal{C}} B_c(\lambda), \alpha_d + \lambda\sigma_d \right\} & \text{if } d \text{ has children and } v_{o\to d} \geq k \\ \alpha_d + \lambda\sigma_d & \text{else} \end{cases}$$

becomes:

$$B_d(\lambda) = \min\left\{ \sum_{c\in\mathcal{C}} B_c(\lambda), (|o| + |d|)v_{o\to d} \right\}$$

where $B_d(\lambda) = \sum_{c\in\mathcal{C}} B_c(\lambda)$ means that the optimal solution implies splitting the d (so $x^* = 1$), and $B_d(\lambda) = (|o| + |d|)v_{o\to d}$ means that the optimal solution implies keeping d aggregated (so $x^* = 0$). So we have to prove Eq. 3.4:

$$\sum_{c\in\mathcal{C}} B_c(\lambda) < (|o| + |d|)v_{o\to d}. \tag{3.4}$$

We separate $\mathcal{C}$ into the set of children with volume above k and the set of children below k:

$$\mathcal{C}^+ = \left\{c \in \mathcal{C}, v_{o\to c} \geq k\right\}$$
$$\mathcal{C}^- = \left\{c \in \mathcal{C}, v_{o\to c} < k\right\}$$

Then, by definition of $B_c(\lambda)$:

$$\sum_{c\in\mathcal{C}} B_c(\lambda) \leq \sum_{c\in\mathcal{C}^+} (|o| + |c|)v_{o\to c} + \lambda \sum_{c\in\mathcal{C}^-} v_{o\to c},$$

and $|d|$ and $v_{o\to d}$ can be expressed as sums over the children:

$$\begin{aligned}(|o| + |d|)v_{o\to d} &= \left(|o| + \sum_{c\in\mathcal{C}^+} |c| + \sum_{c\in\mathcal{C}^-} |c|\right)\left(\sum_{c\in\mathcal{C}^+} v_{o\to c} + \sum_{c\in\mathcal{C}^-} v_{o\to c}\right) \\ &= |o| \sum_{c\in\mathcal{C}^+} v_{o\to c} + \sum_{c\in\mathcal{C}^+} |c| \sum_{c\in\mathcal{C}^+} v_{o\to c} + \sum_{c\in\mathcal{C}^-} |c| \sum_{c\in\mathcal{C}^+} v_{o\to c} + (|o| + |d|) \sum_{c\in\mathcal{C}^-} v_{o\to c}\end{aligned}$$

where:

$$\sum_{c\in\mathcal{C}^+} |c| \sum_{c\in\mathcal{C}^+} v_{o\to c} \geq \sum_{c\in\mathcal{C}^+} |c| v_{o\to c} \qquad \text{(strict inequality if } \mathcal{C}^+ \text{ is not empty)}$$
$$\sum_{c\in\mathcal{C}^-} |c| \sum_{c\in\mathcal{C}^+} v_{o\to c} \geq 0$$
$$(|o| + |d|) \sum_{c\in\mathcal{C}^-} v_{o\to c} \geq \lambda \sum_{c\in\mathcal{C}^-} v_{o\to c} \qquad \text{(strict inequality if } \mathcal{C}^- \text{ is not empty)}$$

It follows that Eq. 3.4 is true:

$$(|o| + |d|)v_{o\to d} > \sum_{c\in\mathcal{C}^+} (|o| + |c|)v_{o\to c} + \lambda \sum_{c\in\mathcal{C}^-} v_{o\to c} \geq \sum_{c\in\mathcal{C}} B_c(\lambda)$$

We proved that for any d with children such that $v_{o\to d} \geq k$, the best pruning error of d is obtained by splitting d. In order to ensure that d is actually split in the optimal solution, we

also have to prove that the precedence constraint is met. If we note a a node on the path from the $root(T)$ to d, then we know that $|a| > |d|$. So $|o| + |a| > |o| + |d| > \lambda$, so the optimal solution also implies splitting a. As it is true for all nodes from $root(T)$ to a, by induction d is split in the optimal solution.

□

3.3.5 Global suppression constraint

Solving problem H or D separately for each origin $o \in \mathcal{O}$ requires distributing the suppression constraint C so that the solutions together do not suppress more than a volume C. Choosing such a distribution is in itself an optimization problem. We can instead solve for all the destinations at once, by considering a tree $\tilde{T}$ made of $|\mathcal{O}|$ times the same hierarchy T, connected together with a dummy root as illustrated in Fig. 3.14. Each subtree represents the destination map of an origin, and thus solving the problem on tree $\tilde{T}$ yields one destination map for each origin in $\mathcal{O}$. Solving the problem for $\tilde{T}$ instead of T is computationally much more demanding, hence the importance of problem D solved with SBA, which scales very well for large trees.

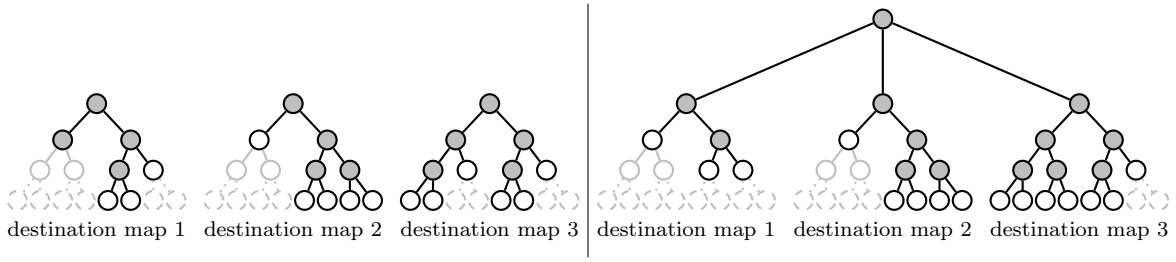


Figure 3.14: Left: separately generalizing the destination map of each origin. Right: solving for the global problem under the unified constraint C

Note that as the soft suppression constraint transforms the constraint into a penalty associated to each node, and not dependent on the nodes outside of its children, there is no suppression constraint to distribute when we solve the soft problem H' for all origins at once.

3.3.6 Proposed models

Based on the problems we formulated, we propose two approaches to k-anonymize OD matrices through generalization and suppression:

- **Adaptative Tree Generalization (ATG)-Dual**: In this approach, we uncouple the anonymization of origins and destinations by first generalizing the map of origins by solving problem P on the Value Generalization Hierarchy T, then solving the dual problem D for destinations with the Specific Breakpoints Algorithm on the tree $\tilde{T}$, composed of one T for each origin linked together by a dummy root.

- **ATG-Soft**: As the solving of the soft problem H' also yields results with interesting properties and very light computing cost, we evaluate it with λ set to 10% of the map size. Note that setting λ to 100% of the map size would be equivalent to using common state-of-the art definitions of the generalization error, which penalize suppressed records

as if they were generalized to the maximum possible level. However, this also leads to prohibitively high suppression costs with solutions that suppress almost nothing. This illustrates that the number of modalities in mobility data is far greater than in other domains, as this problem does not arise in the anonymization of regular data.

3.3.7 Parameter choices

Defining a generalization hierarchy. The hierarchy T is central to our approach. To ensure the best results, it could be valuable to manually define the areas and organise the order in which they ought to be aggregated, in order to give priority to areas that are known to share the same land usage. In this study, we use a satisfying automated approach, using the implicit spatial partitioning of the data when available, *e.g.,* Voronoi tessellation of base stations for mobile phone data. We then run a hierarchical agglomerative clustering on the centroids of the initial areas, and we use the resulting dendrogram as a generalization hierarchy.

Selection of v_{target}. The decoupling of origins and destinations introduces a hyper-parameter v_{target} that needs to be tuned. A simple criterion can be considered in order to choose a value: setting v_{target} too high will lead to origins more coarsely generalized than destinations, and the contrary for v_{target} too low. As the original goal is to produce an all-purpose anonymization, we have no reason to favor precision towards origin or destination. Moreover, we observe through experimentation that the least G is obtained for origins and destinations of the same mean size. It is then recommended to set v_{target} such that the mean sizes of origins and destinations are the closest. The best value depends on the number of areas in the study zone, the total volume of the matrix, and the structure of the flows. In this study, we set a single v_{target} for each data set, selected via a grid-search on a small sample of matrices.

3.4 Data

We evaluate the various approaches on a selection of anonymization problems obtained from open data and from mobile traces. In the following, we detail the pre-processing we performed to get an anonymization problem defined by OD matrices and a Value generalization Hierarchy from sets of spatio-temporal data. Table 3.2 summarises their main characteristics: we detail in particular the **density** of the matrix, computed as the number of non-null flows over the total possible number of distinct flows (*i.e.,* the squared number of tiles), averaged over the period of observation, and the **percentage of flows that are 10-anonymous** in the original data. The relatively low amount of these 10-anonymous flows corresponds to a significant share of the total volumes of the data, as they are naturally the biggest flows. The part of the anonymous volume in the original data is given under the column **%anon_vol**.

New York taxis (nyc) The first problem, which we note `nyc`, consists of all records of yellow and green taxis for 2019 from the New York Taxi and Limousine Commission (TLC). [1]. The data includes the zone of origin and zone of destination for more than 22 million trips over the course of the year. The geography of the 263 tiles is defined by the TLC based on the NYC Department of City Planning's neighborhood definitions. We partition the data set

[1] publicly available at: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

into 7,242 matrices corresponding to one-hour time slices, and each trip is considered in the time step of its departure. The generalization hierarchy defined on these zones is obtained as the dendrogram of a hierarchical clustering applied on the centroids of the zones, with $L2$ distance and Ward linkage. This is the only problem we consider that uses open data, as the others rely on mobile phone data.

Ivory coast (civ) The second problem is noted `civ`, and is based on the mobile trajectories available from the first D4D challenge (Blondel et al., 2013), covering 5M users from 1st December 2011 to 28th April 2012, defined on 1,221 base stations. With the main idea of generating OD matrices without considering their actual ground truth, we consider any transition between antennas as an actual trip. The OD matrices are then defined as the counts of the trips made between antennas that start during a one-hour period. The initial tiles considered for these data sets are the Voronoi tessellation of the base stations projected in `epsg:2165`. The generalization hierarchy we use is the dendrogram of a hierarchical clustering applied to the base stations, with $L2$ distance and station linkage. This approach is also used for the other anonymization problems that use similar data.

Senegal (senegal) This problem uses mobile trajectories available from the second D4D challenge (de Montjoye et al., 2014), covering 9 million mobile users from 7th January to 22nd December 2013, defined on 1,666 base stations. Only 300,000 users are visible in the data at a time, and the batches roll over a two-week period. In the same approach as for the `civ` data set, we consider each transition between base stations as an actual trip, and each OD matrix describes the trips over a one-hour period. The generalization hierarchy is also defined as the dendrogram of a hierarchical clustering with the same parameters as for `civ`. The next anonymization problems are based on OD matrices from the `senegal` problem in order to better explore the behavior of the anonymization methods depending on the characteristics of the OD matrices.

West Senegal (senegal_crop) To understand the impact of the number of initial zones on the computation time, we consider the OD matrices of `senegal` defined only on the 599 westernmost antennas, the most urbanized area of Senegal.

Senegal inflated (senegal_big) The Senegal trajectories cover 300,000 users on a rolling 2-week basis. To generate a heavier data set, we make the hypothesis that people have roughly the same behavior throughout the year, which allows us to aggregate all matrices with the same time step modulo two weeks into a matrix consisting of the trips of 9M users.

Senegal with artificial cells (senegal_split) To study the effects of larger maps, we artificially divide each Senegalese base station into four dummy antennas. Each trip between two antennas is randomly assigned to one of the 16 possible trips between the sub-antennas according to a non-uniform multinomial law: the origin and destination are independently sampled from a law with parameters $(0.4, 0.3, 0.2, 0.1)$, with the sub-antennas ordered. We chose not to assign the trips uniformly so that the best solution would not start by trivially aggregating all sub-antennas to their original antennas, and to maintain a low graph density as would be expected from an OD matrix defined by many domains.

name	#matrices	#tiles	average #flows	density	average volume	%anon. flows	%anon. vol
nyc	7242	263	3058	4.4%	15009	15.8%	59.0%
civ	632	1221	3523	0.2%	3117	0.3%	2.6%
senegal	6752	1666	18276	0.7%	41027	5.7%	36.8%
senegal_crop	6528	599	13710	3.8%	29614	5.8%	37.8%
senegal_split	360	6664	305194	0.7%	956742	6.5%	46.0%
senegal_big	360	1666	100322	3.6%	956742	12.3%	80.7%

Table 3.2: Descriptive statistics of the data sets used for the experiments (#matrices: number of matrices available in the whole data set, where each matrix represents the flows over a time step; #tiles: number of initial tiles over which the matrices are set; density: average graph density of the matrices; avg. #flows: average number of flows among the matrices in the data set; avg. vol.: average sum of the flows; %anon. flows: fraction of flows that are above $k = 10$; %anon. vol: fraction of individuals that are in flows above $k = 10$). Note that due to performance concerns, we do not evaluate the benchmark on all available matrices and we rather choose a small sample of matrices for each data set.

3.5 Anonymization benchmark

In this section, we discuss the approaches that we compare in our benchmark. Except for one approach using a scikit-learn implementation, all the approaches have been re-implemented in python both because a readily usable implementation was not available and to ensure comparable computing times. Some of these approaches have required slight adaptations in order to be applicable to our problem, which we also describe.

Adaptative Tree generalization: We naturally include both our approaches in the benchmark, which both require a given hyper-parameter v_{target} for the decoupling of origins and destinations. For each data set, we select v_{target} based on a grid-search performed over a small training set. The values obtained for various values of k are detailed in table 3.3. Note how the best value for v_{target} is actually not very sensitive of the value of k.

data set	$k=5$	$k=6$	$k=7$	$k=8$	$k=9$	$k=10$	$k=11$	$k=12$	$k=13$	$k=14$	$k=15$
civ	100	100	100	100	100	200	100	200	200	200	200
nyc	100	100	100	100	100	100	100	100	100	100	200
senegal						300					
senegal_big						400					
senegal_crop	200	300	300	300	300	400	400	400	400	400	400
senegal_split						500					

Table 3.3: Best v_target for the studied data sets for various values of k.

Glove: We compare our solution to Glove (Gramaglia and Fiore, 2015), which is the state-of-the-art for k-anonymization of mobility data. We implement Glove in python. The initialization step requires to compute the distance between all couples of records (in our case,

all couples of flows), and at each step Glove merges the two closest records and compute the distance between the result and each other cluster. A significant improvement in time performance in the initialization step can be achieved by only computing the distance between the couples featured in a nearest-neighbors graph. Then, after merging two flows we compute the distance of the result only with the clusters that are a neighbors of one of the two merged flows. We use a 100-nearest neighbors from scikit-learn (Pedregosa et al., 2011) without sensibly impacting the coarseness of its generalization.

By nature, Glove does not consider a generalization hierarchy and takes as input coordinates of the points of the trajectories, performing spatial generalization. For this study, we assign as coordinates the coordinates of the centroids of the tiles of the map. As our approach does not explicitly implement l-diversity and t-closeness, we could not measure the benefits offered by Tu et al. (2017) which would only give a coarser generalization than Glove at a higher computing cost.

Glove-sk: Glove is drastically simplified when applied to simple ODs, seen as 2-point trajectories without timestamps. In that case, it is almost equivalent to the greedy clustering underlying the approach, *i.e.,* to a regular hierarchical agglomerative clustering (HAC) on the four-dimension points defined by the coordinates of origins and destinations, with complete linkage and L_1 metric. As such, we also propose Glove-sk, a simple implementation of Glove using scikit-learn (Pedregosa et al., 2011) making use of the efficient implementation of HAC to get a full dendrogram, and then choosing the lowest cut that grants k-anonymity to all but C records. Unfortunately, because it can only be agnostic of the volumes of the flows, scikit-learn's HAC performs appallingly badly compared to actual Glove. We still include it in the benchmark as it is by far the most readily available solution for a data owner, running in a reasonable amount of time for data that are not as high-volume as the ones featured in this study. As the implementation offered by scikit-learn benefits from state-of-the-art code optimization, it also acts as a lower bound for the computing time we could expect from Glove in the case of the best possible implementation. As in our Glove implementation, we also add a 100-nearest neighbors to speed up calculations, as is recommended by scikit-learn's documentation for hierarchical clustering with high number of points. However, the adaptation does not exactly returns what Glove would, because of two main differences:

- First, the complete linkage considers the distance between two clusters to be the maximum distance between the points of each cluster. Instead of considering the points of the clusters, Glove considers the smallest hypercube containing the points. The distance used by Glove is maximum distance between points of the hypercubes.
- More importantly, the hierarchical clustering is agnostic of the volume and eventually suggests merging flows that are both above the anonymity threshold k. This leads to a huge generalization overhead.

Because of this, the scikit-learn implementation of Glove performs appallingly worse than actual Glove. We still include it in the benchmark as it is by far the most readily available solution for a data owner, running in a reasonable amount of time for data that are not as high-volume as the ones featured in this study. As the implementation offered by scikit-learn benefits from state of the art code optimization, it also acts as a lower bound for the computing

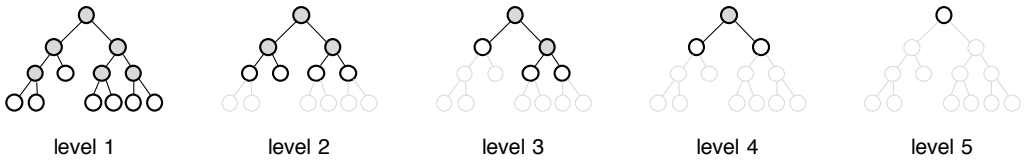


Figure 3.15: All generalization levels considered by OIGH for a hierarchy that does not have all leaves at the same depth.

time we could expect from glove in the case of the best possible implementation.

Mondrian: Although Mondrian has been found to perform very poorly on trajectories (Abul et al., 2010), we observed that it produces relevant results for OD matrices. We compare our approaches to our own python implementation of Mondrian, which performs cuts in the coordinates of the flows seen as 4D points. As it takes coordinates as input instead of qualitative areas, we assign to each area the coordinates of the centroid, as in Glove. At each step, we chose to cut the dimension that has the highest range, and we cut it to the median. If this cut implies separating a cluster into clusters that have less than a volume of k, we do not cut and we stop the algorithm in the current branch. As Mondrian forms a tree, the other branches make their own independent cuts.

OIGH: We also compare our own python implementation of uniform generalization with the OIGH algorithm, which gives horizontal cuts in the hierarchies of origins and destinations. The hierarchies we use, derived from dendrograms, do not have all leaves at the same tree depth, which is originally not handled by OIGH. The OIGH algorithm has been adapted to match the particular structure of the generalization hierarchies of this studies: Uniform generalizations usually consider all the initial values to be at the same tree depth which is not necessarily the case for us. Some steps of OIGH could then yield generalizations that are not partitioning of the study area. In order to keep a coherent generalization at all steps, going up one level in our implementation only generalize nodes that are of minimum-cardinal among the available parents. This problematic is illustrated in Fig. 3.15.

Differential privacy: For comparison, and although differential privacy does not answer our problem, we compare k-anonymization to 1-differential privacy reached through the noise process described in Section 3.2.3, based on laplace noise. The results obtained by **suppression** alone are also shown in the result tables. We implement a variation of the noise process described by Dwork and Roth (2014), that results in integer, non-negative noisy flows: the recommended Laplacian noise is completed by a deterministic rounding and a thresholding to 0 for negative flows. As differential privacy is immune to post-processing (Dwork and Roth, 2014), these operations preserve the ϵ-differential privacy property (Eq. 3.1). This process implies that 30.33% of the flows of volume zero featured in the original OD matrix have a non-zero volume in the anonymized OD matrix, which considering the sparsity results in an anonymized OD-matrix between 6 and 126 times heavier to store. However, it is still better than applying the noise without other processing, which would result in an anonymized OD matrix that almost entirely made of noise and that would be between 20 and 400 times heavier than the original OD matrix, depending on the sparsity.

3.6 Discussion on the performance indicators

For the choice of performance indicators, we bore in mind that the particular generalized areas we provide will most likely not be of interest for the final users of the data. It is reasonable to assume that data users will rather be interested in drawing their own areas of interest and querying the matrix for an estimation of the flows between these areas. This is especially true as the various anonymized OD matrices that represent different timesteps from a single data set will feature generalizations that are *a priori* inconsistent, and comparing the timesteps requires projecting them on a single static zoning. In the absence of additional information, the best estimation flow between two arbitrary areas is obtained by considering the volumes as uniformly distributed in their generalized areas, and as a result the flow between the arbitrary areas is proportional to the overlap with the generalized areas. We call this estimation process **reconstruction**. For more details on the use of the reconstruction process for temporal analysis, see appendix 7.4

It is then relevant to actually reconstruct the anonymized OD matrix and evaluate the difference with the original. It amounts to considering that k-anonymization is obtained by the addition of a kind of reconstruction noise, which makes it a good way of comparing k-anonymization with the laplacian noise used for differential privacy. In the results, we report the **reconstruction loss** E computed as the absolute difference across all flows:

$$E = \frac{1}{V} \sum_{o,d \in leaves(T)} |\tilde{v}_{o\to d} - v_{o\to d}|, \tag{3.5}$$

where we denote $\tilde{v}_{o\to d}$ the reconstructed volume over these initial tiles, and we normalize by the total volume of the flows $V = \sum_{o,d\in leaves(T)} v_{o\to d}$ for readability. Note that the sum is over the leaves of T, corresponding to the initial tiles over which the OD matrix is defined.

It is important to note the behavior of the reconstruction loss, to better interpret its value as in indicator for the information loss of an anonymization algorithm. OD matrices are by nature sparse and they can also have sparse emission and reception maps, for example if the matrix is set over a fine grid. The k-anonymized version of such an OD matrix features generalized origins with a small number of emission hot spots separated by a high number of empty tiles, and similarly for destinations. For the sake of simplicity, we consider a generalized flow $o \to d$ containing two hot spots each contributing $v_{o\to d}/2$, and a number z_o and z_d of empty tiles in o and d, respectively. The reconstructed flows uniformly distribute $v_{o\to d}$ into all $z_o \times z_d$ possible flows. This means a difference of $\frac{v_{o\to d}}{z_o z_d}$ for the empty flows, and $\frac{v_{o\to d}}{2} - \frac{v_{o\to d}}{2 z_o z_d}$ for the two hotspots. We obtain the absolute difference:

$$\begin{aligned} E_{o\to d} &= (z_o z_d - 2)\frac{v_{o\to d}}{z_o z_d} + v_{o\to d} - \frac{v_{o\to d}}{z_o z_d} \\ &= 2v_{o\to d} - 3\frac{v_{o\to d}}{z_o z_d} \end{aligned}$$

which quickly converge to $2v_{o\to d}$ when $z_o z_d$ grows. As long as $z_o \approx z_d \geq 15$, the noise then mostly represents the volume of the flow, regardless of its other properties, which is not a desirable feature.

We also evaluate the approaches with respect to their total generalization error G, as

defined in Eq. G. Approaches such as Glove and Mondrian that do not rely on a generalization hierarchy do not have a clearly defined value for G. As they use an axis-aligned bounding box as a reference for their clusters, we count all the tiles intersecting with the box in the size of the generalized area. We give here a general expression for G that is more readily applicable to all generalization methods, defined over any set $\mathcal{F}^+$ of anonymous flows $o \to d$ such that $v_{o\to d} \geq k$. The value we report in the results is the **mean generalization error** $\bar{G}$ across matrices, given by:

$$\bar{G} = \frac{1}{V^+} \sum_{o\to d \in \mathcal{F}^+} (|o| + |d|) v_{o\to d}, \tag{3.6}$$

with $V^+ = \sum_{o\to d\in\mathcal{F}^+} v_{o\to d}$ the total volume of anonymized flows. When the origins and destinations are aggregated to roughly the same level as is normally the case, $\bar{G}$ represents roughly twice the number of tiles in the origin or destination of the average generalized flow. A value of $\bar{G} = 2$ then means that no generalization was performed.

Finally, as a complementary metric to measure the distortion induced by the anonymization, we evaluate the **total variation** L_1 between each OD matrix and its anonymized versions. This metric is slightly different than the reconstruction loss E as it considers OD matrices as normalised distributions, not penalising suppression. The distribution distance is given by:

$$L_1 = \sum_{o,d\in leaves(T)} \left| \frac{\tilde{v}_{o\to d}}{V^+} - \frac{v_{o\to d}}{V} \right| \tag{3.7}$$

with the same notation as the previous definitions.

3.7 Results

We compare the approaches on a restricted number of OD matrices on the available data sets. For each data set, the matrices were chosen at random among matrices with a total volume of more than 5 000. For each matrix, we set the suppression constraint C to be 10% of the total volume and we set $k = 10$ in accordance with the value accepted by the French regulator CNIL for OD matrices. Fig 3.16 gives a data set by data set comparison of our ATG-Dual approach versus ATG-Soft, OIGH, Glove, and Glove-sk on metrics $\bar{G}$, E, distribution, and time of computing. For each data set, we represent the distribution of the performance of the benchmark solver, expressed as the ratio with the performance of ATG-Dual. The line $y = 1$ indicates matrices for which ATG-Dual and the benchmark have the same value, and the line $y = 10$ indicates matrices for which the competitor's metric is 10 times higher than ATG-Dual. As it is best to have a low error and a low computing time, ATG-Dual is better when the distributions are above the line $y = 1$.

Unsurprisingly, we see that ATG-Soft is faster but coarser than ATG-Dual, as it is essentially a cheaper version that uses a fixed value for λ instead of finding the best one. ATG-Dual gives consistently better results than OIGH in a shorter computing time.

Upon closer inspection, we observe that OIGH is likely to waste time in its search inside the lattice of possibilities, because it yields a lot of ties that then need to be broken. Glove offers a noticeably finer generalisation than ATG-Dual. However, its prohibitive computing time makes it difficult to recommend, and its memory usage is such that we were unable

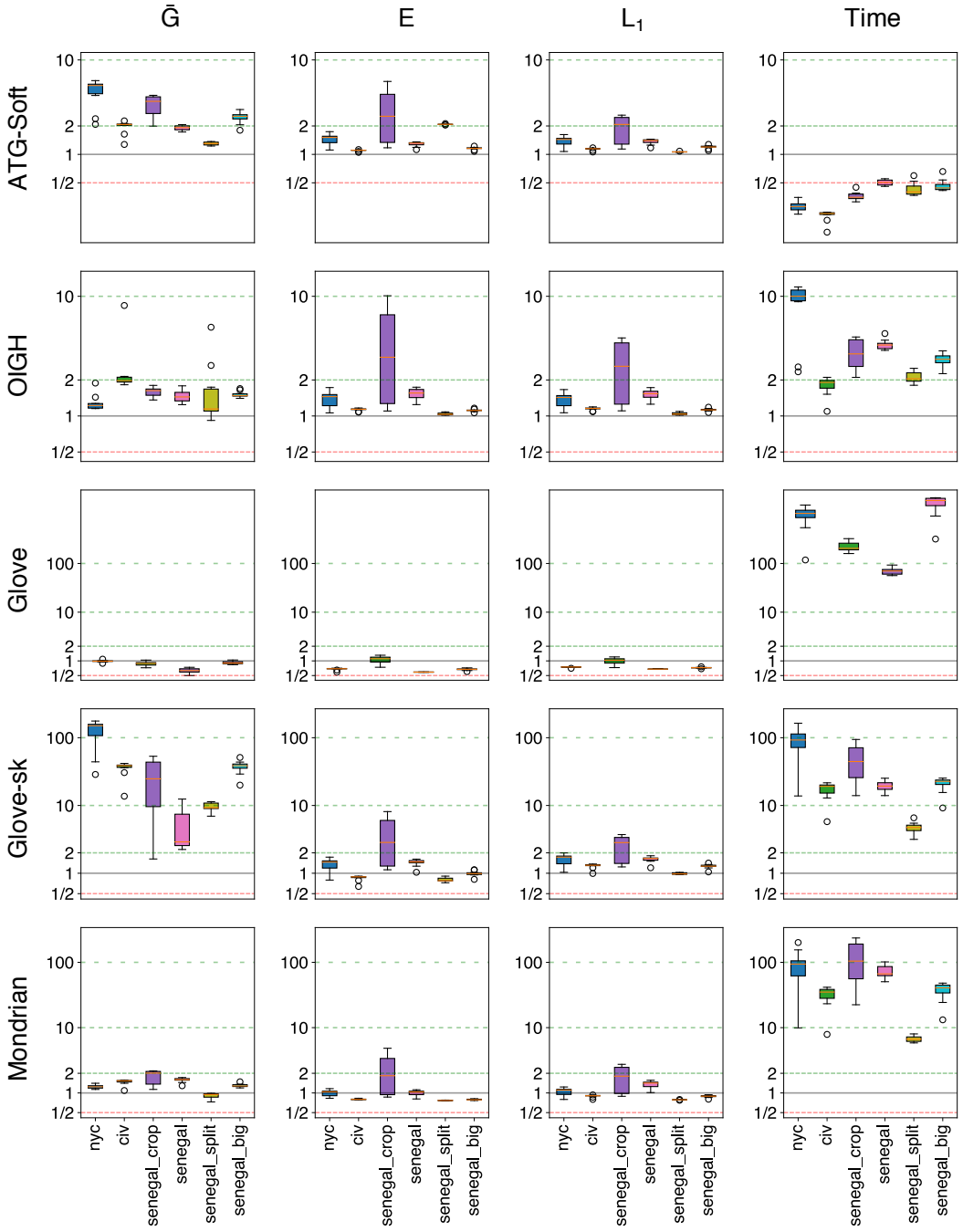


Figure 3.16: Comparison of the performance of our ATG-Dual approach versus ATG-Soft, Glove-sk, Glove, OIGH, and Mondrian. Each box represents the distribution of the performance over one data set, expressed as the ratio of the benchmark over ATG-Dual. First column: comparing generalisation error G. Second column: comparing the reconstruction loss E. Third column: comparing the distribution distance L_1. Fourth column: comparing the computing times. In each case, the line $y = 1$ represents matrices for which the benchmark has the same performance as ATG-Dual, and the lines $y = p$ represents matrices for which the benchmark is p times worse than ATG-Dual. The box plots correspond to the data sets in this order: `nyc`, `civ`, `senegal_crop`, `senegal`, `senegal_split`, `senegal_big`.

to run it for our bigger data sets `senegal_split` and `senegal_big`. An optimistic take on Glove would be to consider that it has the performance of Glove with the computing time of Glove-sk. Glove-sk in itself offers remarkably bad results, which shows the importance of the volume awareness in the hierarchical clustering proposed by Glove, that could not be handled with HAC. We include it in the comparison as it gives an estimation of the computing time that a truly optimized implementation of Glove could possibly offer. Yet, its computing time is still unsatisfactory compared to our approaches. Mondrian finds generalisations that are coarser than our approach for all data sets except `civ`, but the reconstruction error is much more balanced, with an advantage for Mondrian. This is at least partly explained by the fact that Mondrian does not suppress volume, which has a direct impact on the absolute errors measured in E: Each unit of missing volume has a contribution of 1 in the expression of E (Eq. 3.5), which means that an additional 1% of suppression contributes to an increase of 0.01 in E. As Mondrian requires repeatedly counting volumes on both sides of the cuts we consider, it is much slower than ATG-Dual. Some implementation improvement could be considered in order to optimize the counting steps, but it is unlikely that it could achieve a 100-fold reduction in computing time.

Table 3.4 summarises the average performance over all matrices in the small data sets `nyc`, `civ`, `senegal` and `senegal_crop`, for which we were able to run Glove. We report the mean value of $\bar{G}$, E, L_1 and S as well as the total computing time to anonymise all the matrices. We see that Glove stands out for its lack of scalability, making it impractical in huge-volume cases. Among the other solutions, a difference of minutes that we measure in the computing time is relevant but not decisive. They must rather be compared based on $\bar{G}$, for which ATG-Dual offers a significant improvement over the state of the art. On the other metrics, Mondrian performs remarkably well: it has a better reconstruction loss E, as noted previously due to the fact that it does not suppress volumes, but also offers a better L_1. An explanation for this performance would be that Mondrian happens to find generalisations that are more uniform in terms of volumes. For differential privacy, E and S mostly measure the volumes that have been added to 0-flows, as they represent more than 95% of possible flows in each OD matrix. Because of this, differential privacy performs worse than any generalisation technique when compared on E. This matches the previous observation (Cormode et al., 2011) that differential privacy is not adapted for sparse data. As the interpretation of $\bar{G}$ is only relevant for generalisation, we do not include it in the table for suppression and differential privacy.

As Glove could not be evaluated for our biggest data sets, we compare the approaches on the same average criterias in a separate Table 3.5 for `senegal_split` and `senegal_big`. We see that the best computing time we could hope for Glove, given by Glove-sk, is still not satisfactory in this situation. Our method is of particular interest here, as it offers a finer generalisation than OIGH and Mondrian for a fraction of the time. The ATG-Soft alternative is even faster, yet the difference in computing time at this scale is not of importance. This approach could become relevant for even bigger matrices, for example set on the ten thousand base stations of a mobile operator in France.

Even if our approach is more appropriate, uniform tree generalisation admittedly performs well for OD matrix generalisation. Indeed, we could expect the best solution to show a high disparity of aggregation levels between densely and sparsely populated areas, which a uniform generalisation cannot offer. This effect is mitigated by the fact that the initial tiles already

solver	$\bar{G}$	E	L_1	time (s)	S
ATG-Dual	24.90	1.18	1.14	23	9.99%
ATG-Soft	40.41	1.76	1.31	9	3.24%
OIGH	53.62	1.34	1.30	56	9.87%
Glove-sk	444.12	1.10	1.36	358	9.99%
Glove	18.70	0.81	0.84	23203	9.98%
Mondrian	27.39	0.95	1.02	688	0.00%
Suppression	—	0.33	1.34	—	39.46%
Differential privacy	—	2.88	6.78	—	-170.92%

Table 3.4: Performance on samples of the data set senegal_crop, nyc, civ, and senegal. $\bar{G}$: mean generalisation error (Eq. 3.6); E: normalized reconstruction loss (Eq. 3.5); S: fraction of volumes suppressed (Eq. S); L_1: distribution distance (Eq. 3.7). The reported time is the total computing time (in seconds) to run the anonymisation of all matrices of the samples. Note that differential privacy adds volumes, as it mostly applies a positive noise on a sparse matrix.

solver	$\bar{G}$	E	L_1	time (s)	S
ATG-Dual	**19.85**	**0.72**	**0.78**	111	9.92%
ATG-Soft	65.39	1.07	1.09	**32**	1.54%
OIGH	29.68	1.10	1.15	985	9.79%
Glove-sk	1361.51	1.04	1.25	3376	10.00%
Mondrian	25.19	0.76	0.85	11541	0.00%
Suppression	—	0.49	1.02	—	59.00%
Differential privacy	—	2.87	3.33	—	-86.98%

Table 3.5: Performance on samples of the data sets senegal_big and senegal_split.

partially reflect the disparity in activity density, as they rely on base stations or administrative divisions, which are more densely distributed in populated areas. This illustrates the importance of the generalisation hierarchies, as well as the adaptation we implemented in order for OIGH to run on hierarchies whose initial leaves are not all at the same depth. For the detailed table of results by data set and solver, see Appendix 7.3.

3.8 Conclusion on the anonymization

In this chapter, we proposed to k-anonymize OD matrices that are large-scale both in terms of size of the study area and number of flows. To that end, we developed **ATG-Dual**, a tree-based approach that formulates generalization and suppression as an optimization problem and finds the best adaptative generalization, as opposed to a uniform generalization usually found by similar tree-based approaches. For even heavier OD matrices, we also propose **ATG-Soft**, a faster version of ATG-Dual that relies on a fixed parameter λ instead of finding

the best one, sacrificing precision of generalization for the benefit of computing speed. The formulation we adopt is indeed a restriction on the solution space we consider. State-of-the-art solutions developed for mobility consider broader search spaces as they allow any form of clustering of flows. This is a natural way of considering k-anonymity, and allowing more degrees of freedom is essential in the context of trajectory anonymization, as they are especially hard to anonymize. Yet even with these techniques it has been observed that the valuable information cannot be expected to resist k-anonymization for $k \geq 5$. By considering OD matrices, which are a central input to mobility analysis while being considerably simpler than trajectories, the data can be anonymized with good preservation of information. The generalization hierarchy then appears as a relevant restriction that is natural to the data user and considerably reduces the computing cost for huge matrices, where traditional approaches struggle to scale. Our two approaches find the best anonymization in their respective solution spaces under a hard constraint: ATG-Dual enforces a hard constraint on the suppression, which is especially relevant in the context of OD matrices with a high number of modalities, as the representativity of the data is essential to their value. ATG-Soft enforces a hard constraint on the size of generalization, which is also of interest to guarantee the precision of the data. Uniform tree generalization, while being rather scalable and offering acceptable results given the right adaptations, is still slower and coarser than our approaches for huge OD matrices. In particular, as it was originally designed to generalize numerous attributes with very small hierarchies of less than a dozen modalities, it is only natural to focus on a finer cut in the hierarchy when we have only two attributes with thousands of modalities. Our approaches come at the cost of decoupling the attributes through an *ad hoc* process requiring a hyper-parameter v_{target}. Although in our current solution it has to be set by hand based on the results over a small subset of the data, it should be possible in future work to infer the best value for v_{target} based on characteristics of the input OD matrix. The cost function used for the generalization of origins could also in itself be modified in order to better anticipate the processing of destinations.

Our work has the practical aim of helping the owners of large-scale OD matrices, such as mobile operators, to efficiently and effortlessly anonymize their data. However, it is well known that k-anonymity does not protect against all types of privacy attack. The approach should guarantee l-diversity in order to ensure protection against attribute linkage attacks, meaning the generalized areas should all cover at least l locations. We can consider as location either the initial areas in themselves, or the various points of interest contained in each initial area. In each case, this condition could be guaranteed in future work by adding a minimum level of aggregation as a constraint. In the broader scope of probabilistic attacks, future work should focus on achieving t-closeness, meaning the distributions of destinations given each origin should not differ too much from the global distribution of destinations, and likewise for the distributions of origins. As the criterion requires to know the whole destination distribution, it is not local: we cannot introduce an additional term in each node of the best pruning problem and try to minimize its sum. The variation of the problem to ensure t-closeness would take the form of an additional constraint akin to the suppression constraint, and likely require relaxation as well in order to efficiently solve it. Still, t-closeness would remain hard to achieve in the context of sparse distributions with high number of modalities such as OD matrices. It may also be of interest to measure the actual privacy risks of the OD matrices in terms of probability of success of various attack scenarios. Indeed, OD matrices are already arguably

safer than most types of personal data as the number of attributes is very limited.

The k-anonymity offered by our approach is thus a necessary first step toward a more complete solution. Still, it is also a sufficient guarantee in itself for the French regulator CNIL, in charge of enforcing GDPR in France. Currently, mobility data tend to be under-used by their owners as their huge volumes and their personal aspect make their handling costly and legally risky. Achieving cheap, fast, and foolproof anonymization of mobility data would allow their widespread public use, ensuring the most insights possible are extracted from them in order to organise ever more efficient transportation networks.

Chapter 4

Synthetic population with activity chains

4.1 Introduction

The dynamics of urban transportation can be captured using activity-based models, which rely on synthetic travel demand data to get a comprehensive understanding of urban mobility. For example, mobility simulators such as Matsim (Horni et al., 2016) can give a dynamic overview of the travel behaviors of a population provided it is given as input a representative and realistic population of agents with their agendas, their constraints, and their preferences. This information is usually derived from a collection of sources rather that one single comprehensive data set. Micro-samples of population often serve as the basis to generate a complete synthetic population of agents described by their socio-economic factors (Rasouli and Timmermans, 2013). Depending on the selected approach, aggregated statistics can also be used as a reference for calibration, for example, to ensure the right age or income distribution in the synthetic population (Saadi et al., 2018; Sun et al., 2018). Work and study locations can be determined with the help of commute matrices (Hörl and Balac, 2021; Fournier et al., 2021), while the data for secondary locations comes from passive sources such as mobile phones (Bonnetain et al., 2021). The problem of generating such a **synthetic travel demand** can be divided into three parts:

- Generating the socio-economic features of the agents;
- Assigning activity chains to the agents;
- Setting locations for the primary and the secondary activities.

Extensive work has been proposed to generate complete synthetic travel demand or parts of it. The socio-economic features alone can be modeled by probabilistic graph models (Sun and Erath, 2015; Sun et al., 2018) or deterministic rescaling approaches (Yameogo et al., 2021). The primary locations can be considered independently from the activity chains, in the sense that we expect all agents to have a home location, and depending on their status they can also have a work location or a study location, regardless of their actual agenda for the day. Consequently, some synthetic population approaches such as Rich (2018) or Chapuis et al.

(2018) generate the primary locations jointly with the socio-economic factors. The locations can be drawn during the generation of the activity chains when using generative approaches such as the one proposed by Song et al. (2010) or Anda et al. (2021). Alternatively, the activity chains can be generated first without locations (Axhausen and Herz, 1989; Joubert and De Waal, 2020), while trajectories, without the semantic information of the activity chains, can be generated separately (Lin et al., 2017; Pappalardo and Simini, 2018; Ouyang et al., 2018). The activities then have to been inferred from the traces with approaches such as Jiang et al. (2017) or Yin et al. (2018). Note that to our knowledge, no approaches exist yet to jointly generate socio-economic factors, activity chains, and locations in a single model. Doing so would give the benefit of a unified generative framework for the complete travel demand, although the approach would certainly suffer from having to generate such a high-dimensionality distribution.

In this chapter, we delve in more detail into the various approaches used for the first two parts of the synthetic travel demand, that is the socio-economic factors and the activity chains. We then present the data that we use in this chapter and in Chapter 5 for our own synthetic travel demand that follows the principle of the state-of-the-art pipeline proposed by Hörl and Balac (2021). We describe our approach to calibrate the synthetic population with respect to the temporal distribution of trips observed in mobile data. The spatialization of the population will be handled in Chapter 5.

4.2 Generating a synthetic population

4.2.1 Socio-economic features

Generating socio-economic features for a set of synthetic persons is usually referred to as the problem of **synthetic population**. It relies on a **population micro-sample** describing a restricted number of persons at the individual level, and marginal statistics over the whole population, such as the total number of male, female, elderly, office workers, *etc.* In this problem, the micro-sample is usually very detailed (containing the full description of the individuals) but very small (accounting for less than 5% of the actual population). It usually comes from dedicated surveys designed to gather more advanced insights that what is possible with aggregated census results. In comparison, the marginal statistics are very coarse, as they only give an information on the total number of people with one specific value for one specific variable (such as the number of people for which the gender is female), without any information on the relations between the variables. As basic, official descriptions of the country's population, they are however considered the most reliable information we can have about a population. They are usually derived from the official census, which acts as a reference for all such usages.

The problem is then to create a full population made of individuals whose socio-economic descriptions are realistic compared to the micro-sample, and such that the marginal totals also fit the ones observed by the census.

Note that although the population is composed of individuals, several population synthesis approaches focus on the ability to generate a **hierarchical population** made of individuals grouped into households. This distinction can be useful for attributes that are directly related

to the household, such as the number of cars or the type of habitation. Counting the totals of these variables by summing over the individuals would not give the right result, as the cars and the homes are shared inside a same household. More importantly, individuals from a same household are expected to display very correlated activity chains (Bhat et al., 1999; Anggraini, 2009): if a child needs to be accompanied to each of their activities, then each of their trips will correspond to a trip in the agenda of their parents. Conversely, if one person in the household handles groceries, then the others probably won't need to do it. However, although hierarchical population has been explored in the context of the synthesis of socio-economic variables, approaches taking it into account for the generation of activity chains are yet to be developed to our knowledge.

Iterative Proportional Fitting (IPF): IPF is the first approach to population synthesis, popularized with Beckman et al. (1996) and still occasionally in use today due to its satisfying performance. It is applicable when all features of the population are categorical and the total number of people for each value of each feature is known, forming a tensor where each entry is the number of people for one value, as illustrated in Fig. 4.1. It consists in applying a rescaling factor to all the people sharing a same value of a feature so that the target is matched. As long as we rescale different values of a same feature, no contradiction appears since we rescale different individuals. As we start rescaling based on another feature, the resulting population match the target totals for the second feature but does not match the first feature anymore. However, it has been proven than the process eventually converges (Sinkhorn, 1964) and that the limit is the maximum likelihood estimator (Bishop, 1967). Due to its simplicity, its solid mathematical justification, and its good practical performance, IPF has long been the most prominent approach for all rescaling problems and is still occasionally in use in recent work (Saadi et al., 2018).

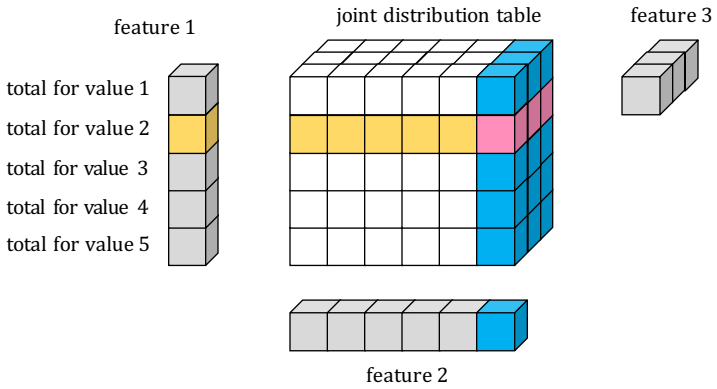


Figure 4.1: IPF setting for a population with 3 features. The population is represented by a 3-dimensional tensor where each entry is the number of people for a given joint value. The target totals are given by vectors. For any given feature, its target total vector can be matched by applying the appropriate scaling factors to each slice (for example the yellow slice for the second value of feature 1). As we proceed to calibrate the population with respect to feature 2, the scaling factors associated to the pink cells can be modified, but the algorithm eventually converges.

Iterative Proportional Updating (IPU): IPF is unable to account for hierarchical population, specifically populations made of agents organized in households. IPU (Ye et al., 2009; Konduri et al., 2016) has been introduced as a variant to tackle this specific problem. IPU applies in a different context than IPF in two aspects: First, the agents that we rescale (households) are separated into sub-agents (individuals), and aggregated statistics are used as target totals for both the number of agents and sub-agents. Second, the aggregated statistics are not marginals (like "number of households that have a car") but are rather totals of specific types of agents, defined by the joint value of all their attributes. The functioning of IPU is illustrated in Table. 4.1. In this example, 23 people are divided among 8 households. Each person can be either of type P1, P2, or P3 (defined by a joint value of socio-economic variables), while each household is either of type H1 or H2 (for example, defined by their location or the type of dwelling). We assume that the actual total number of each of these types is known, which is given in the line "Target total". The first iteration assigns scaling factors to all households of type H1 such that the total is matched. The scaling factors are shown in the columns on the right, while the corresponding lines give the corresponding totals of the population taking into account those scaling factors. Rescaling the population so that the number of households of type H1 matches the target naturally does not impact the number of households of type H2, but impact the number of persons of each type, as some of them are in households of type H1. After having rescaled with respect to all the household types, the totals for the persons have changed but does not necessarily match the targets. Iteration number 3 updates the scaling factor of all the households that contain a person of type P1, with the result that the marginal for P1 is perfectly matched but not the household marginals anymore. After a pass through each type, we see on the bottom line that the population totals are closer but not perfectly matched to the targets. In practice, multiple iterations of this process converge to the desired output.

	Household type		Person type			Weight iteration					
	H1	H2	P1	P2	P3	0	1	2	3	4	5
	1		1	1	1	1	11,67	11,67	9,51	8,05	12,37
	1		1	0	1	1	11,67	11,67	9,51	9,51	14,61
	1		2	1	0	1	11,67	11,6	9,51	8,0	8,05
		1	1	0	2	1	1,00	13,0	10,59	10,5	16,28
		1	0	2	1	1	1,00	13,00	13,00	11,0	16,91
		1	1	1	0	1	1,00	13,00	10,59	8,9	8,97
		1	2	1	2	1	1,00	13,00	10,59	8,9	13,78
		1	1	1	0	1	1,00	13,00	10,59	8,9	8,97
Target total	**35**	**65**	**91**	**65**	**104**						
Weighted sum 0	3,00	5,00	9,00	7,00	7,00						
Weighted sum 1	**35,00**	5,00	51,67	28,33	28,33						
Weighted sum 2	35,00	**65,00**	111,67	88,33	88,33						
Weighted sum 3	28,52	55,38	**91,00**	76,80	74,39						
Weighted sum 4	25,60	48,50	80,11	**65,00**	67,68						
Weighted sum 5	35,02	64,90	104,84	85,94	**104,00**						

Table 4.1: Example of IPU to rescale a population of households containing persons. Example from Ye et al. (2009).

Other notable approaches for rescaling a hierarchical population include an entropy optimization method (Bar-Gera et al., 2009), and Hierarchical Iterative Proportional Fitting (HIPF) (Müller and Axhausen, 2010), which applies scaling factors to both households and individuals.

As an optimization problem: Since the problem amounts to assigning weights to the rows of a matrix in order to match a given target, the synthetic population problem can also naturally be formulated as an optimization problem (Fournier et al., 2021; Ye et al., 2020). A first formulation of such a problem would be of the form:

$$\begin{aligned} &\min_x \|Ax - b\| \\ &\text{s.t. } x \geq 0 \end{aligned} \tag{4.1}$$

where A is the data matrix representing the households and their compositions, x is the vector of rescaling factors, and b is the vector of targets for each joint value. Note that in this case each dimension in b corresponds to the number of a given household type or person type, as in IPU, and not to a marginal count as in IPF. Note also that because of the matrix multiplication, the households must be represented as columns in the data matrix A. Using the same example as in Tab. 4.1, we can detail the structure of the matrix in the following equation, where we give the interpretation of each row of the matrix:

$$Ax - b = \begin{matrix} H1 \\ H2 \\ P1 \\ P2 \\ P3 \end{matrix} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 1 & 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 0 & 2 & 1 & 1 & 1 \\ 1 & 1 & 0 & 2 & 1 & 0 & 2 & 0 \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix} - \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix} \tag{4.2}$$

As we can see, the problem is severely underconstrained: The number of constraints is the size of the target vector, which is given by the number of person types plus the number of household types, which in typical cases would give a few hundred rows. Meanwhile, associating one rescaling coefficient to each households can easily lead to tens of thousands of coefficients. This problem can be mitigated by grouping equivalent households that share the same household type and the same composition of persons. In that case, each group of equivalence contain n_i households and the rescaling coefficient x_i found for group i must be equally distributed among the n_i household in the group. However, even with this simplification step, the number of degrees of freedom must be expected to be much higher than the number of marginals, that act as constraints.

Generalized raking: From the point of view of surveys, this formulation as an optimization problem is an opportunity to penalize and control the distortion of the population induced

by the rescaling. It has been studied under the name of calibration problem, where the main motivation is to make a survey sample match a given composition(Deville et al., 1993; Davies, 2018). For example, for a survey containing 40% men and 60% women, 30% students and 70% non-students, what would be the result if the survey had instead contained 50% of each category. The calibration problem looks for the rescaling coefficients that are the closest to 1 while respecting the targets:

$$\min_{\substack{x \geq 0 \\ Ax = b}} \Phi(x) \tag{4.3}$$

where Φ is an arbitrary function of the form $\Phi(x) = \sum_j \phi(x_j)$, ϕ itself being an arbitrary function that is is minimum in 1. Three main functions are used for ϕ:

- The quadratic function: $\phi(x_j) = \frac{1}{2}(x_j - 1)^2$
- The raking function: $\phi(x_j) = x_j \ln(x_j) - x_j + 1$
- The logit function: $\phi(x_j) = \frac{(1-l)(u-1)}{u-l}\left((x_j - l)\ln\left(\frac{g-l}{1-l}\right) + (u - x_j)\ln\left(\frac{u-x_j}{u-1}\right)\right)$, where l and u are lower and upper bounds set on the rescaling coefficient, respectively. When using this function, problem 4.3 has the additional constraint $l \leq x \leq u$.

A comparison of these three function is illustrated in Fig. 4.2. The problem is solved by considering the lagrangian obtained by relaxing the equality constraint $Ax = b$:

$$L(x, \lambda) = \Phi(x) - \lambda(Ax - b) \tag{4.4}$$

Looking for x_j such that $\frac{\partial L}{\partial x_j} = 0$ yields an equation on λ that can be solved using the Newton-Raphson method. Then, the solution x^* to the primal problem that is associated with the dual solution λ^* can be easily derived and is feasible (Davies, 2018).

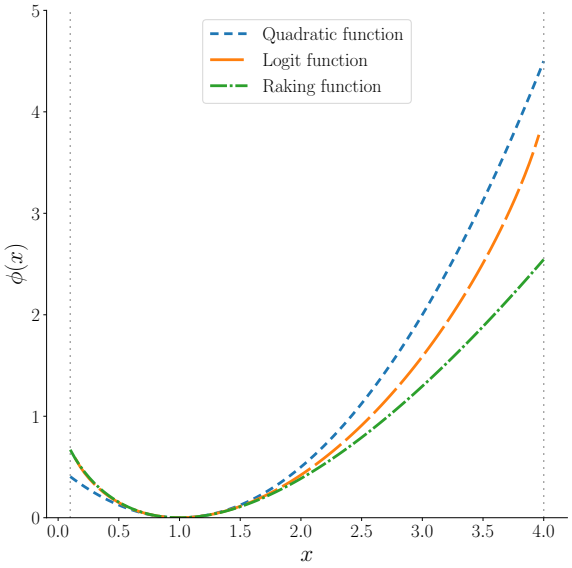


Figure 4.2: Illustration of the quadratic function, the raking function, and the logit function used for the calibration problem. The bounds are $(l, u) = (0.1, 5)$.

The **deterministic approaches** described above are limited to associating a rescaling coefficient to a small set of pre-existing individuals, which means that they can only duplicate

individuals and not actually generate a population. This is a problem when the input micro-sample is very small or the number of variables is high, as it leads to numbers of people for each joint value that are too small to be considered reliable estimates. In the most extreme cases, it leads to situations known as the zero-cell problem (Müller and Axhausen, 2010), where a no individual exist in the micro-sample with a given joint socio-economic value, and none can be created by mere rescaling. This calls for advanced models capable of modelling the underlying probability law in order to generalize from small samples.

Stochastic approaches: When the micro-sample is very small, **stochastic approaches** have been developed that aim to model the joint probability distribution of the features observed in the micro-sample. This then allows to draw a population of any size, potentially creating individuals that do not exist in the micro-sample. Probabilistic graph models are widely used for the synthetic population problem: Sun and Erath (2015) learn the structure of a Bayesian network *via* tabu search to represent the dependencies of seven socio-economic variables of the population of Singapore. For the synthesis of a whole household, they introduce additional nodes in the network of the individuals, representing the status of the spouse, that make it possible to condition the profile of an individual on the profile of their spouse. Later, Sun et al. (2018) use a multilevel latent class model to explicitly describe the household composition, illustrated in Fig. 4.3. Each person is supposed to belong to a latent class, which separately defines a probability law for each of their socio-economic attribute. The households are also separated into latent classes, which affect their sizes and define the parameters of the latent classes of their individuals. In this model, each class of household has their own set of latent classes for the individuals. A simpler variation of the model defines universal classes, where the latent classes of individuals are defined in a first step, and the household class only affect the probability of their people being of a given class. As the model features latent variables, the learning of its parameters must be done with an Expectation-Maximization (EM) approach.

This model suffers from the fact that the profiles of individuals only interact through the household class node, and thus are independent of each other given the household. This is a rather weak representation of the strong interactions that exist between people in the same household. A very simple consequence is, for example, that the model will very often generate same-sex couples, which cannot be representative of the population of Singapore, as same-sex relationships between men were still against the law at the time of the survey used for the input data (2012). The distribution of household structures is reproduced using rejection sampling. However, rejection sampling requires knowledge of the joint distribution of a selection of variables, which is precisely the problem at hand. Thus, the rejection rule can only rely on a criterion based on a very small number of variables. They give as an example the distribution of age and gender difference for two-person households, but a different rejection rule must be designed for each household structure.

Other Bayesian networks and their special cases of Hidden Markov Models have also been proposed for population synthesis applied to Belgium (Saadi et al., 2016), Jakarta (Ilahi and Axhausen, 2019), or Cape Town (Joubert and De Waal, 2020).

Such approaches are however not guaranteed to produce the right marginal totals like the deterministic approaches, which are a common way of validating a synthetic population.

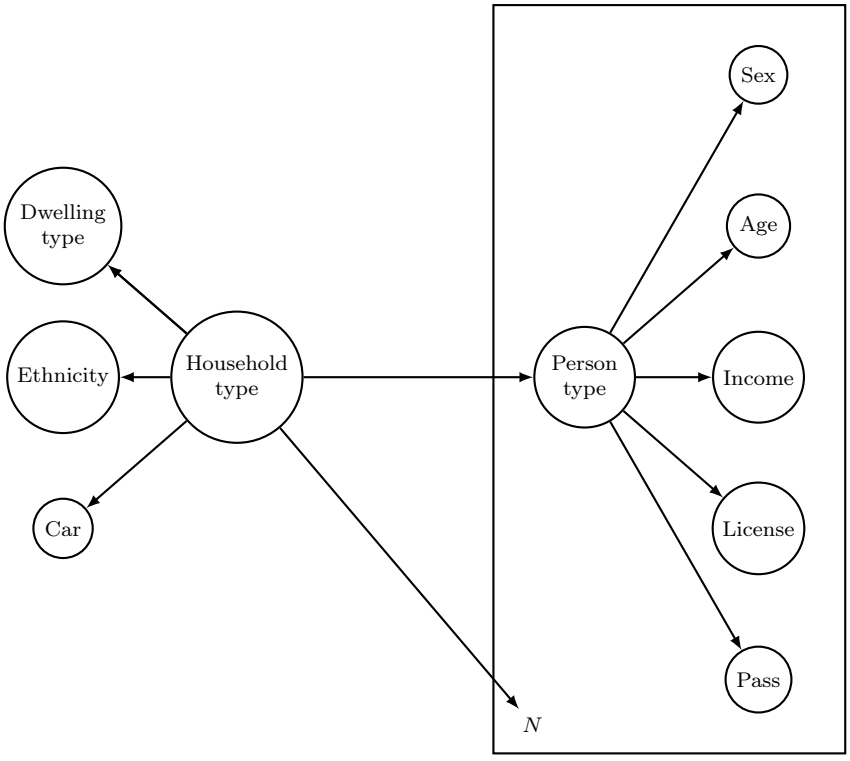


Figure 4.3: Hierarchical model proposed by Sun et al. (2018) The plate notation indicates that a random number N of persons are independently drawn once the household type is decided.

Recent works then propose to combine the result of the graph model, which is efficient to derive a diverse population from a very restricted input micro-sample, with a deterministic rescaling approach such as IPF or generalized raking to ensure that the marginals of the population fit the target measured in a census (Saadi et al., 2018; Ilahi and Axhausen, 2019). More recently, neural networks, especially variational autoencoders, have been used with the aim to offer a better scalability with respect to the number of attributes in situations where the micro-sample is very small (Borysov et al., 2019; Garrido et al., 2020; Johnsen et al., 2022; Aemmer and MacKenzie, 2022). For a more thorough review of popular methods with a focus on hierarchical populations, see (Yameogo et al., 2020).

4.2.2 Integerization

Note that the deterministic approaches yield non-integer rescaling coefficient. While some models do not consider this to be a problem (Rich, 2018), it can hamper the interpretation and the use of the resulting population. To achieve integer coefficients, Ballas et al. (2005) compared several deterministic approaches: Starting from the the straightforward rounding method that tend to induce severe variations in the population size, they then propose an approach to approximate a rounding threshold that minimizes this variation, as well as another approach that cascades the decimal part of the coefficients down to the next individual. All of these determinisitic approaches are found to be outperformed by probabilistic approaches in terms of final population counts and accuracy (Lovelace et al., 2015):

- The first one, called the **proportional probabilities** approach, draws a target number of agents with probabilities proportional to their rescaling coefficient.

- The second one, called the **Truncate, Replicate, Sample (TRS)** method, or stochastic rounding (Gupta et al., 2015)), rounds up each coefficient with probability equal to its decimal part, and rounds it down otherwise.

These two methods are also compared in Yameogo et al. (2021), who finds that TRS outperforms proportional probabilities in terms of matching to the original population.

4.2.3 Activity chain assignment

Activity chain assignment is the problem of providing each agent, defined at this point only by socio-economic factors, with a chain of activities defined by their purposes, the time of the day, and optionally the transport mode between each activity. The location of each activity is not a part of this problem, and is handled in a further step in the generation of synthetic demand. The activity chains can either be generated or taken from an HTS.

Generative approaches rely on probabilistic graphical models: Works such as Liu et al. (2015); Saadi et al. (2016) propose a Hidden Markov Model to generate the sequence of activity types, with a distinct model for different profiles identified in the population. As the approaches focus on **tours**, *i.e.,* activity chains that start and end at home, forward sampling is impossible and so the models are sampled from using Markov Chain Monte Carlo methods. Joubert and De Waal (2020) propose a Bayesian network that both handle the socio-economic features of the population and the agenda, with two variables for each trip (one for the transport mode and one for the activity purpose). This integrated approach allows to condition on variables to study the impacts on the others, which can be valuable: for example, it can yield an estimation of how many trips by car are performed by someone who has a driver license, or give a different distribution for the agenda depending on whether the person is employed or not. An expanded model taking the duration of the activities is also proposed. The structure of the network is learned using tabu search, and conveniently describes a chain where each activity is only dependent on the previous one, and the first activity is dependent on the socio-economic profile. This seemingly satisfactory structure hides the fact that the agendas being of variable lengths, the activities can be `None`, which is encoded like a regular value. This can be justified by considering that it is not a missing information, but instead an information that there is no i^{th} activity in the agenda. However, we can argue that it is mostly a consequence of using a relational data format with a fixed number of columns, which has a decisive impact on the learning of the structure: as a `None` activity necessarily implies that the next activity is also `None`, the algorithm finds a very strong correlation between an activity and the next, although it can actually be more of an artefact of the data encoding.

When an HTS is available, some works (He et al., 2020; Hörl and Balac, 2021) use an approach based on the concept of **statistical matching** (D'Orazio et al., 2006). This process is close to a regular join between relational databases with a join key of multiple variables, except that each agent is required to match a minimum number of HTS respondents. For agents that do not, the variables are dropped from the join key in a predefined order, one at a time, until the agent match the required number of HTS respondents. One of these HTS rows is then drawn at random, and its activity chain is assigned to the synthetic agent. The locations given in the HTS row are not used and must be generated in the next steps. An example is given in Table 4.2 and Table 4.3. The first table (Table 4.2) contains a single row

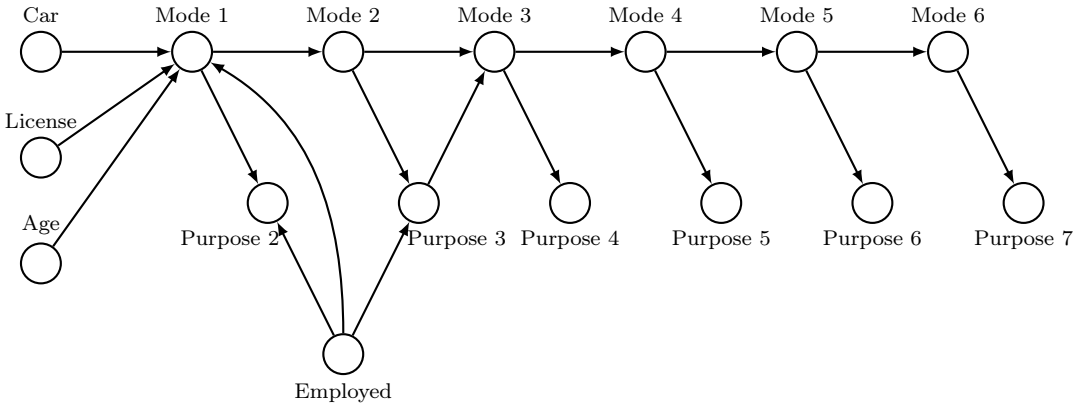


Figure 4.4: Bayesian network proposed by Joubert and De Waal (2020) to model both the population and the agendas, learned by tabu search.

representing an agent of the synthetic population to which we want to assign an activity chain from the HTS. The second table (Table 4.3) represents an HTS, where each respondent has an activity chain. If we perform a relational join between the two tables based on all features we have at hand, we obtain zero match for our agent. If we remove the last column, *i.e.,* `Home` status|, then the relational join gives us two matches. In this example, if we want at least three matches, we need to consider only the first three variables `age`, `gender` and `occupation`. This gives us three potential activity chains, among which we select one at random.

Agent	Age	Gender	Occupation	Car availability	Home status
1	30-59	female	executive	yes	tenant

Table 4.2: Example of a synthetic agent requiring an activity chain.

4.2.4 Evaluating a synthetic population

The evaluation of synthetic demand is made difficult by the high dimensionality of the data and the lack of ground truth source. The only available data, that is population micro-samples and marginal distributions over the total population, are used as input to the model. This leaves only **internal validation**, *i.e.,* comparing the synthetic population with the original input datasets. For the evaluation of the socio-economic features, the difference between the joint distribution of the synthetic population and the input micro-sample can be measured as a coefficient of determination, a total absolute error, a standardized error, a modified Z-score, or a Root Mean Squared Error (RMSE) (Lovelace et al., 2015). The common practice is to evaluate the fit of the synthetic population against the known marginal distributions (Anda et al., 2021; Fournier et al., 2021; Hörl and Balac, 2021). Some studies take into account the variability due to the stochastic processes (Rich, 2018; Hörl and Balac, 2021). To that end, several experiments are run to make sure the standard deviation of the marginal composition of the population is small. Studies aiming at generating activity chains perform evaluation based on various chosen indicators: number of trips, exploration sequences, start time of activities, duration of activities, and spatial imprecision (Anda et al., 2021). Joubert and De Waal (2020) only generate activity chains and transport modes, and evaluate the synthetic

HTS respondent	Age	Gender	Occupation	Car availability	Home status
1	30-59	female	executive	yes	owner
2	18-29	male	student	no	tenant
3	30-59	male	executive	yes	tenant
4	18-29	female	employee	no	tenant
5	60+	male	retired	no	owner
6	30-59	female	executive	yes	owner
7	30-59	female	inactive	yes	owner
8	60+	male	independent	yes	tenant
#matches	4	3	2	2	0

Table 4.3: Example of HTS respondents. The matching attributes are highlighted. The last row indicates the number of match that the agent in Tab. 4.2 have when considering the n left-most attributes. For example, if we consider only `age` and `gender`, then there are 3 matches.

activity chains as a dependant variable to infer from socio-economic variables, considering the confusion matrix of their model.

4.2.5 Positioning

In this chapter, we describe how we synthesized our own population for the city of Lyon, France. Most comprehensive state-of-the art approaches rely on a pipeline of steps to take into account each source of information that is at hand. We describe our data in Section 4.3 and detail the steps inspired from the state-of-the-art pipeline proposed by Hörl and Balac (2021). Then, we observe that in the absence of calibration, the temporal distribution of the trips made by our population does not match what is observed from our mobile data. In Section 4.4, we propose a step that can be added to the synthetic population pipeline to calibrate this temporal distribution to the target distribution that we have at hand. The positioning of both the state-of-the-art steps and our contribution, that are described in this chapter, is illustrated in Fig. 4.5. We evaluate our synthetic population in Sec. 4.5 in terms of realism of the socio-economic composition, the temporal distribution, and the popularity of agendas. We conclude in Sec. 4.6. Note that in this chapter, we do not handle the spatialization, which will be discussed in Chapter 5

4.3 Data used for the synthesis

In this section, we describe the study zones, the OD matrices and the surveys used in our work. We also describe the preprocessing that we perform in order to obtain compatible descriptions: as the OD matrices describe three months of mobile phone movements and the HTS describe the activity chains for a typical day, care must be taken both in the extraction of a "typical day" in the OD matrices and in the correspondence between mobile phone movements and trips as defined in the activity chains. We then describe the state-of-the art steps that we performed to obtain a synthetic population for the city of Lyon.

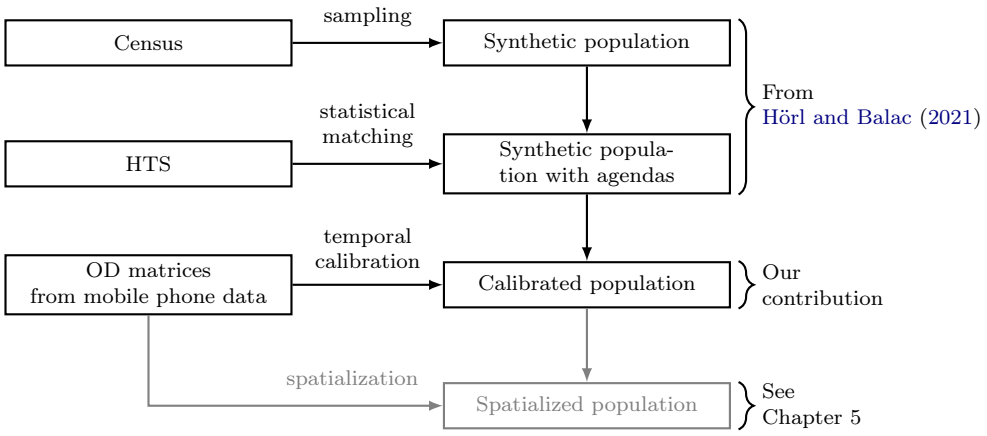


Figure 4.5: Positioning of the population synthesis pipeline and the steps discussed in this chapter.

4.3.1 Zoning

Our study area takes the form of a square 25,600 meters wide centered around Lyon, France, divided into 515 distinct zones which represent either municipalities or spatial units (IRIS) common to many statistical analyses in France. Fig. 4.6 shows the partitioning of the study area.

4.3.2 OD matrices from mobile data

We use OD matrices from (Bonnetain et al., 2021) describing the movement of residents of the region of Lyon, based on trajectory data from French telecommunication operator Orange that have been rescaled to represent the whole resident population. Mobile phone data are defined at the cell level but the work from TRANSIT, re-distributing locations inside the territory of each cell, allows to convert the spatial data to the zoning defined in Fig. 4.6. These data cover three months from 19/03/19 to 18/06/19 and contain information about 49,681,883 distinct flows, that we aggregate to obtain a collection of OD matrices for a typical working day divided in 15 time steps. Note that the week-ends have been removed, as well as the following holidays: 22/04/2019 (Easter Monday), 01/05/2019 (labor Day), 08/05/2019 (1945 armistice), 30/05/2019 (Ascension Day), 31/05/2019 (commonly taken holiday), and 10/06/2019 (Whit Monday). The restriction to our study area gives the flows between the zones performed by resident of the larger Lyon region. The data does not count trips performed by non-residents nor trips whose origin or destination is outside the study area. The mobile phone trajectories are converted to location chains and then OD matrices with the processing described in Sec. 2.2.3.

The resulting OD matrices are interpreted as transition matrices, defining for each couple of zones (o, d), and each time step t of a typical working day the transition probability $P(d|o, t)$. The transition matrices are noticeably less sparse than what is usually expected from OD matrices, due to summing all equivalent time steps over the 59 distinct days. The density of the transition matrix with respect to the time step is detailed in Table 4.4.

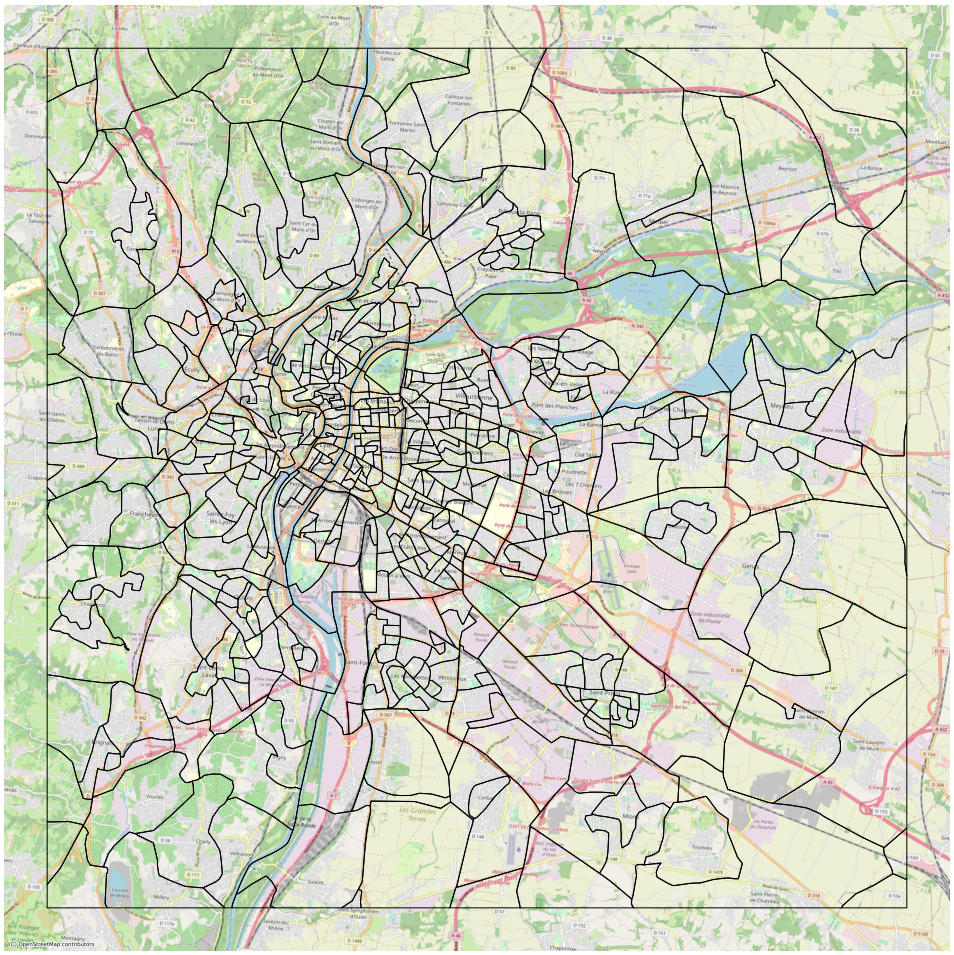


Figure 4.6: Map of the partitioning of the study area of the city of Lyon. Background: OpenStreetMap

4.3.3 National census

The French national census (INSEE, 2018) is a rich dataset containing socio-economic variables for around one third of French residents. Each of the 487,628 rows for our study zone features socioeconomic variables and a scaling coefficient for a total of 1,366,072 persons, described by the variables detailed in Table 4.5. Note that we choose not to retain the household structures of the population. In a generative approach, this information would be necessary for the downstream modeling steps. In our approach, activity chains are assigned based on individual socio-economic attributes, hence only individuals are relevant.

Census is given by IRIS, a fine statistical partitioning. When the number of people in an IRIS is too small to guarantee statistical secrecy, the IRIS is censored and only the canton is given, which is unacceptably coarse. In those cases, we use the similar data sets "Mobpro" and "Mobsco" made available to researchers upon request by French organization Quetelet-Progedo-Diffusion. Mobpro and mobsco contain micro-data similar to the census, the main difference being that the home zones are given at the commune level.

Note that although INSEE widely uses the concept of socio-professional category (INSEE, 2003), we use a slightly modified concept of **occupation** that also has a category for students.

Time step	Transition matrix density
0h-2h	28%
2h-5h	20%
5h-7h	46%
7h-8h	56%
8h-9h	57%
9h-0h	52%
10h-12h	62%
12h-14h	65%
14h-16h	65%
16h-17h	59%
17h-18h	60%
18h-19h	59%
19h-20h	55%
20h-22h	54%
22h-0h	48%

Table 4.4: Density of the transition matrices with respect to the time step.

Variable	Values
Home zone	515 zones in the study area
Age	0-17; 18-29; 30-59; 60+
Gender	Male or female
Occupation	Inactive; farmer; independent; executive; employee; intermediate professions; worker; student; retired
Car availability	Yes or no
Home status	Owner; tenant; social housing

Table 4.5: National census variables used for the synthetic population.

4.3.4 Household Travel Survey (HTS)

The activity chains are taken from the 2015 Enquête Ménage Déplacements performed by the French agency for urban planning (Cerema, 2015), which details the same features, as well as the complete agendas of 28,230 persons in the region of Lyon among which 3,101 are actually empty. An agenda is represented as a list of trips, defined by the features detailed in Table 4.6. Note that the day is divided into time steps of one or more hours to be consistent with the division used for anonymization in the mobile data. The activity chains contained in the HTS feature up to 12 trips during the day.

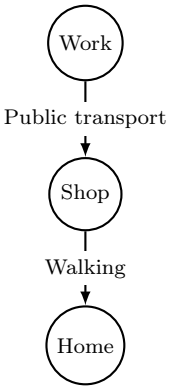
The definitions of trips and activities in a travel survey are different from what is derived

Variable	Values
Activity purpose	Home; work; study; shopping; personal.
Trip time step	0h-2h; 2h-5h; 5h-7h; 7h; 8h; 9h; 10h-12h; 12h-14h; 14h-16h; 16h; 17h; 18h; 19h; 20h-22h; 22h-0h.
Trip transport mode	Foot; bicycle; car or motorcycle; public transport.

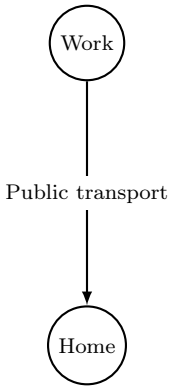
Table 4.6: HTS trip variables describing the activity chains

from the notion of stay points in mobile data: In a survey, any amount of time spent doing something meaningful to the agent counts as an activity, and any movement counts as a trip, even if the destination is the same as the origin (for example when walking a dog). In mobile data, we can only identify areas where a user has stayed for a significant amount of time, which are labeled as stay points, and we consider as trips the segments between consecutive stay points. Since we can only measure the time spent in an area and not how the activity there was actually meaningful to the user, we must assume that important activities last while meaningless stops during a trip don't. This prevents us from detecting short activities in the mobile data. Likewise, the inherent noise in the mobile data prevent us to detect short trips.

HTS pre-processing: To make both data sources compatible, we perform a pre-processing of HTS agendas designed to keep only the trips that would be detected in the mobile data. We delete activities from the agendas in order to ensure that no activity last less than 20 minutes and no trips is shorter than 100 meters. We do so while keeping a meaningful interpretation of the agenda. For example, as illustrated in Fig. 4.7, if someone would leave work using public transport, with only a short stop at a shop before walking home, the short activity should be discarded but the transport mode to go home should be "public transport" and not "walking". The logic of the filtering of the agendas is detailed in Appendix 7.6. Finally, we assume that each agent has a unique location for each primary activity. Consequently, after this first filtering we merge all consecutive home activities, and the same for work and study.



(a) Original agenda



(b) First possible filtered agenda



(c) Second possible filtered agenda

Figure 4.7: Example of merging a short activity (Shopping) with the next one. The least important activity is discarded. If the inbound trip is the longest of the two, then it is considered as the trip going from work to home (b). Else, the outbound trip is considered (c).

The impact of the pre-processing on some characteristics of the data is given in Table 4.7. We see that the number of activities per person is much lower, which is explained by the fact that 1% of the trips are under 100m, and 20% lead to an activity lasting less than 20 minutes. Our preprocessing does not necessarily suppress such trips, as it can merge them instead, but the resulting chain is still one trip shorter. Although we do not suppress many trips based on their length, short activities tend not to require long trips, so the preprocessing leads to overall longer trips. The mean trip length is very sensitive as such mobility data often feature extreme outliers. The median, more stable, show a difference of 500m, which is an acceptable precision for city planning. The distribution of activity purposes is rather stable, while the fraction

of trips performed on foot drops noticeably, for the same reason that short activities tend to be small errands at short, walkable distances. This illustrates that our pre-processing mostly tends to under-represent short trips with a soft transportation mode, which are only a minor problematic in the context of city planning. However, more comprehensive travel demand synthesis would benefit from using alternative sources to model such soft mobility. In fact, a general perspective for future work, which we discuss in Sec. 6.2.7, is to characterize with more precision which trips are well represented by mobile data, and which are under-represented or undetected.

Metric	Original HTS	Filtered HTS
Mean trip length	4196	5285
Median trip length	1220	1739
Mean number of activities by person	4.42	2.75
Median number of activities by person	4	3
Fraction of empty agendas (only one activity)	13%	26%
Fraction of home activities	53%	46%
Fraction of work activities	11%	12%
Fraction of study activities	7%	8%
Fraction of shopping activities	9%	10%
Fraction of other activities	20%	23%
Fraction of trips on foot	32%	25%
Fraction of trips by bike	2%	1%
Fraction of trips by car	52%	57%
Fraction of trips by public transport	15%	17%

Table 4.7: Impact of the preprocessing on the HTS

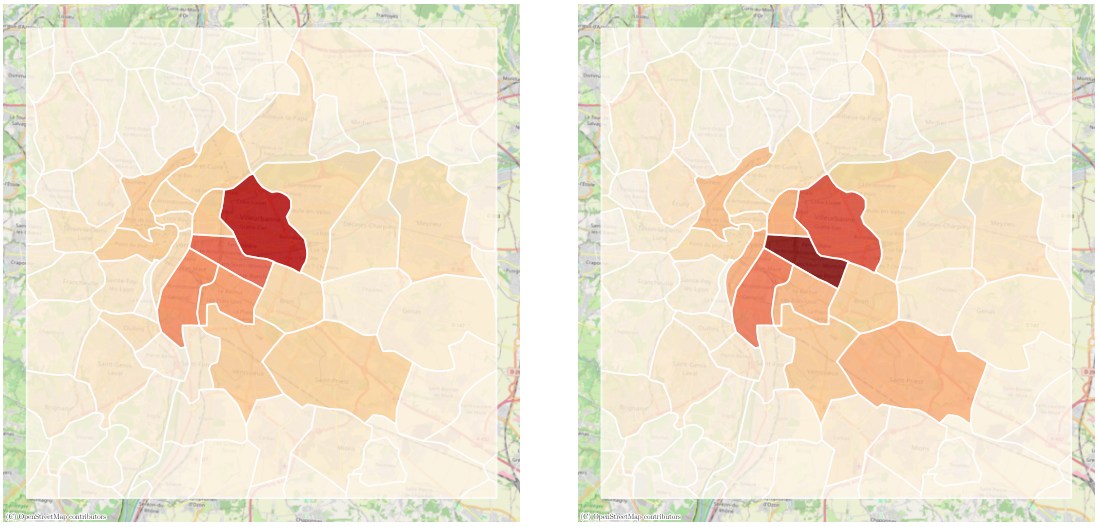
4.3.5 Commute matrices

We use the commute matrix from INSEE as validation data (INSEE, 2020). It is derived from the census and reports the commune of home and the commune of work for a total of 565,816 commuters in our study zone. The commute matrix serves as an indicator for the spatialization to measure how well our approach reproduces the commute patterns of the population, but we also use it to assess the realism of the peak-hour volumes we observed in our other sources. The emission and reception maps of the commute matrix are illustrated in Fig. 4.8, while the matrix in itself is illustrated in Fig. 4.9.

4.3.6 Synthetic population with activity chains

As in Hörl and Balac (2021), the synthetic population is extracted from the census made available by the French statistical institute INSEE, and we perform a statistical matching (D'Orazio et al., 2006) to assign activity chains from the HTS to individuals from the synthetic population. This process is akin to a regular join on a key of multiple attributes between two relational databases, with the following differences:

- We require each row in the synthetic population to match at least $M = 20$ rows of the HTS.



(a) Commute emissions (Home communes) (b) Commute receptions (Work communes)

Figure 4.8: Emission and reception map of the commute matrix for Lyon, France.

- The attributes in the key are ranked by order of importance and the least important attributes can be removed from the key in order to find more matches when necessary.
- Once each row of the synthetic population is matched to a group of a least M rows of the HTS, one of those HTS rows is assigned at random to the synthetic agent, giving them its agenda. In our case, as weights are available for the HTS rows, we perform a weighted drawing.

Statistical matching aims at ensuring diversity in the assignment of chains to synthetic agents while joining as many variables as possible so that the agents have activity chains relevant to their characteristics. The variables used in the join key are described in Table 4.8. The resulting population is a set of agents, each with an activity chain specifying their purposes, time of day, and transport modes.

Variable	Number of values	% of agents with more than 20 matches
Age	4	100%
Gender	2	100%
Occupation	9	95%
Has car	2	94%
Home status	3	90%
Canton	8	75%

Table 4.8: Variables used in the statistical matching to assign activity chains from HTS to agents from census

This results in a synthetic population defined by four socio-economic variables and a home location. This state-of-the-art pipeline uses only data that is at least available for free upon demand for research: the official statistical zoning used by the French statistical institute, the national census, and the HTS. Next, we observe that the trips made by the resulting population

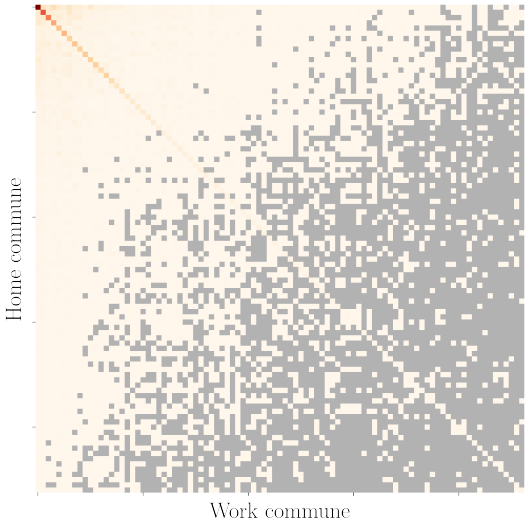


Figure 4.9: Commute matrix of Lyon.

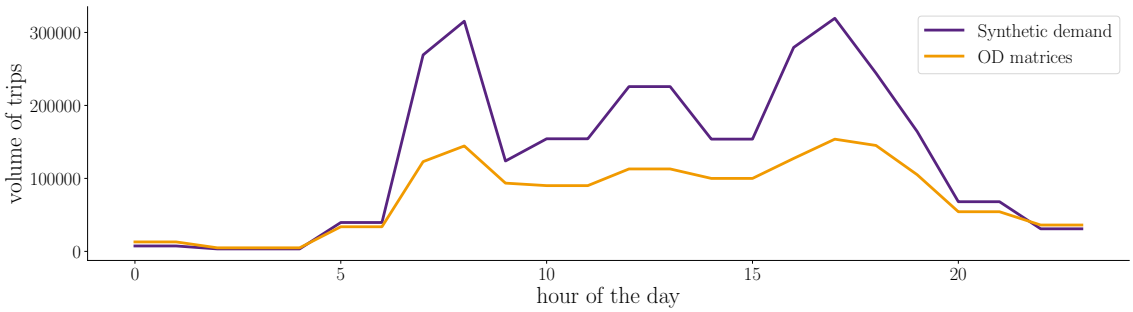
do not have the same distribution during the day than what is observed in the mobile data. We thus propose a step to ensure this calibration, which preserves the socio-economic composition of the population.

4.4 Temporal calibration

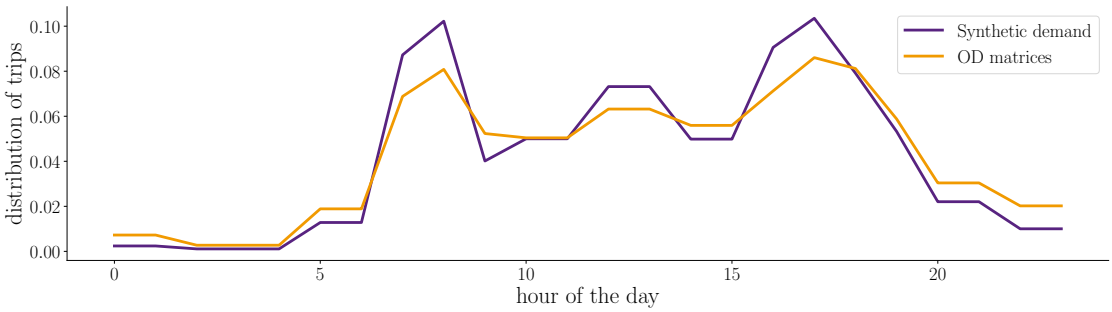
As a result of the operations described in Sec. 4.3, we obtain a population of synthetic agents defined by socio-economic features and an activity chain. The activity chain of each agent specifies, among other things, the start and end time of each trip. The average distribution of the start time of the trips during the day is also available from the mobile phone OD matrices. The expected profile of such a distribution is flat during the night, with a high peak in the morning, a slightly lower, wider peak in the afternoon, and a small peak at noon between the two (Spurr et al., 2014). We see in Fig. 4.10 that both distributions have this profile, albeit with some discrepancies. We also see in Fig. 4.10a a difference in scale: during the day, the number of trips registered in the OD matrices is 1,787,219 while the synthetic population performs 3,085,661 trips in the same period. The two distributions are compared in Fig. 4.10b, where the time steps of more than an hour have been uniformly distributed in order to visualize the usual peak hours.

Several factors can explain these discrepancies. The mobile data may detect false trips during the day, flattening the peaks by comparison, or people who don't have commute trips may use their phone more than people who do. Errors may also come from the HTS, which features a low response rate (CNIS, 2018) possibly harming the estimations.

As an optimistic stand, we assume that the true distribution is given by the mobile data, which were rescaled to correct the customer bias. Our motivation is that the widespread use of this source is bound to improve their quality in the future, and the ability to generate recent mobile data on short notice makes them an ideal candidate for updating expensive, difficult-to-produce HTS. Because an HTS represents only a small number of people and suffers from a very high non-response rate, we might also expect a lack of representativeness for fine-grained



(a) Hourly number of trips performed by our synthetic population vs. number of trips observed from the mobile data.



(b) Temporal distribution of trips performed by our synthetic population vs. observed from the mobile data.

Figure 4.10: Comparison of the trips performed by the synthetic population and the mobile phone data, in actual volumes and in distribution.

statistics. In particular, surveys have been observed to over-represent peak-hour trips and under-represent off-peak demand (Spurr et al., 2014).

However, the actual volumes reported by mobile data are known to be more suitably interpreted up to a scaling factor (Mamei et al., 2019). In particular, the official commute matrix for this area reports at total of 522,647 commuters, which is more compatible with the volumes observed in our synthetic population than in the OD matrices.

Therefore, we propose to calibrate the synthetic population to the distribution of trips observed from mobile data, while maintaining the total number of trips made throughout the day. The socioeconomic distribution of the population must also be respected. To this end, we assign individual rescaling coefficient to the synthetic agents.

4.4.1 Problem statement

In this section and the next, we make heavy use of notations to formalize the calibration problem. All notations are summarized in Table 4.9. The temporal calibration problem can be seen as a hierarchical population problem, except that instead of households composed of individuals, we rescale individuals composed of trips. More formally, we must find rescaling coefficients $c_i \geq 0$ for each of the N individual of the population such that the temporal constraint and the social constraint are satisfied. The trips are described only by their departure time step $t \in \{1, \dots, T\}$. For each i, we note $\hat{v}_{it} \in \mathbb{N}$ the number of trips performed by indi-

vidual i at time step t. It can be 0, 1, or more than 1 as the time steps are between 1 and 3 hours long. We note $v_t^* \in \mathbb{N}$ the number of trips observed at time step t in the mobile data. $\hat{V}$ and V^* denote the total number of trips during the day performed by the population before rescaling and observed in the mobile data, respectively. Then, calibrating the distribution of the trips during the day is formalized as respecting the constraint:

$$\forall t, \frac{1}{\hat{V}} \sum_{i=1}^{N} \hat{v}_{it} c_i = \frac{1}{V^*} v_t^*. \tag{4.5}$$

The individuals are described by four variables: `gender`, `car availability`, `occupation`, and `age category`, for a total of $K = 127$ distinct descriptions observed in the population. For each possible joint value k of these variables, we note $\hat{s}_{ik} = 1$ if individual i is described by k and $\hat{s}_{ik} = 0$ otherwise. Then, the rescaling respecting the distribution of the population is formalized as respecting the constraint:

$$\forall 1 \leq k \leq K, \sum_{i=1}^{N} \hat{s}_{ik} c_i = \sum_{i=1}^{N} \hat{s}_{ik}, \tag{4.6}$$

We compare two approaches to solve this hierarchical population problem. The first is Iterative Proportional Updating (IPU), which is specifically designed for such problems. However, it lacks formal guarantees both in terms of convergence and of properties of the solution found. We compare it to a more formal optimization formulation inspired from Fournier et al. (2021), of the form $\min_{c \geq 0}^{c} ||Ac - b||$ where A is the population matrix and b is the targets.

In both cases, we reduce the dimensionality of the problem by grouping the agents that are equivalent, *i.e.,* agents that have the same socio-economic profile and the same number of trips per time step. We obtain $N = 16,002$ distinct groups, and the indices i denote categories of equivalent individuals instead of the individuals in themselves. This means that compared to the descriptions given above, the values of $\hat{v}_{it}$ can be very high, and $\hat{s}_{ik}$ can be more than 1. The size of the target vector is given by the sum of the number of joint socio-economic values observed in the data and the number of time steps, for a total of $D = 142$ dimensions.

Once the scaling coefficients for each group are determined by either of the approaches, we must then assign individual coefficients to each agent in the groups. We do that by giving an equal share of the group coefficient to each individual. Finally, the scaling coefficients are made integer via the Truncate, Replicate, Sample (TRS) approach (Yameogo et al., 2021), which rounds the coefficients up with probability equal to their decimal part.

4.4.2 Formalization as an optimization problem

As noted above, this optimization problem is severely under-constrained. In such a situation, it is best to formalize the problem with a regularization term in the form of a prior distribution of the rescaling coefficients. In this section, we develop a simple formulation inspired by the works of Ye et al. (2020); Fournier et al. (2021).

The prior distribution of the coefficient is assumed normal, centered around the coefficients obtained by assuming each individual has a coefficient of 1 (that is, the coefficient of the ith category of indistinguishable individuals is centered around its size n_i).

$$p(c_i) \sim \mathcal{N}(n_i, \sigma_c^2), \tag{4.7}$$

where the standard deviation σ_c is a hyper-parameter. The prior distribution of the vector of coefficients is then:

$$p(c) = \prod_{i=1}^{N} \frac{1}{\sigma_c\sqrt{2\pi}} e^{\frac{-(c_i-n_i)^2}{2\sigma_c^2}}, \tag{4.8}$$

where N is the number of groups. This *a priori* distribution of the coefficients must be coupled to a noise distribution, interpreted as the inherent noise explaining the discrepancies between the totals observed in the population (*i.e.,* the vector $A \cdot c$) and the target b. We assume that this noise is normal centered around the the totals of the population:

$$p(b_j|c) \sim \mathcal{N}(A_{j:} \cdot c, \sigma_{b_j}) \tag{4.9}$$

Where we note $A_{j:}$ the jth row of the matrix A, thus $A_{j:} \cdot c$ is the total count for the jth type (household or person) as per the population. We note $\mu_j = A_{j:} \cdot c$.

The noise distribution is then:

$$p(b|c) = \prod_{j=1}^{D} \frac{1}{\sigma_{b_j}\sqrt{2\pi}} e^{\frac{-(b_j-\mu_j)^2}{2\sigma_{b_j}^2}} \tag{4.10}$$

where D is the number of targets. The Bayes Rule gives us the expression of the likelihood $p(c|b)$:

$$p(c|b) = \frac{p(b|c) \times p(c)}{p(b)}, \tag{4.11}$$

where the factor $p(b)$ does not depend on the parameters c, meaning that the likelihood is maximal when $p(b|c) \times p(c)$ is maximal. The problem of finding the best coefficients is then:

$$\max_{c \geq 0} \prod_{j=1}^{D} \frac{1}{\sigma_{b_j}\sqrt{2\pi}} e^{\frac{-(b_j-\mu_j)^2}{2\sigma_{b_j}^2}} \times \prod_{i=1}^{N} \frac{1}{\sigma_c\sqrt{2\pi}} e^{\frac{-(c_i-n_i)^2}{2\sigma_c^2}} \tag{4.12}$$

Maximizing the function is equivalent to minimizing $-\log$ of the function, which gives:

$$\min_{c \geq 0} \sum_{j=1}^{D} \frac{(b_j-\mu_j)^2}{2\sigma_{b_j}^2} + \sum_{i=1}^{N} \frac{-(c_i-n_i)^2}{2\sigma_c^2} \tag{4.13}$$

which is a norm minimization problem with L_2 regularization:

$$\min_{c \geq 0} \alpha ||Ac - b||_2^2 + \beta ||c - n||_2^2 \tag{4.14}$$

where n is the vector $(n_i)_{1 \leq n \leq N}$. This minimization problem is then solved with Adam optimizer (Kingma and Ba, 2014), which is a widely used version of the gradient descent algorithm. The constraint $c \geq 0$ is implemented by projecting the gradient on the constrained space at each step, which is guaranteed to find the minimum of a convex function on a convex subspace (Calamai and Moré, 1987).

The assumption that the coefficient follows *a priori* a normal distribution supposes that there is additive noise in the estimation of the coefficient. This is disputable as these coefficient are not measured from a noisy sensor but are computed based on the survey plan. However it acts as a convenient quadratic regularization. Other assumptions on the *a priori* law could include the log-normal distribution, which is indicated when the noise is supposed multiplicative and naturally implies that the coefficients are non negative. However our experimental results with a log-normal a priori law were not conclusive. The noise distribution, describing the probability of the target given the rescaling coefficients, supposes that the official, target results b are actually noisy estimates of the true totals, which are given by the rescaled population. Here, supposing that this noise is Gaussian makes intuitive sense as the target b comes from a measure of the population, which is likely to contain many small errors adding to each other.

Symbol	Interpretation	Value
t	Denotes a time step of the day.	$t \in \{1, \dots, T\}$
T	Number of time steps in the day.	$T = 15$
k	Denotes a joint value of the socio-economic variables.	$k \in \{1, \dots, K\}$
K	Number of joint socio-economic values observed in the synthetic population before rescaling.	$K = 127$
i	Denotes a category of agents defined by a common agenda and socio-economic description.	$i \in \{1, \dots, N\}$
N	Number of distinct agents, as defined by their agenda and socio-economic description.	$N = 16,002$
n_i	Number of agents belonging to category i.	$n_i \in \mathbb{N}$
n	Vector of counts for each socio-economic category.	$n = (n_i)_{1 \leq i \leq K}$
D	Number of constraints of the problem.	$D = K + T = 142$
V^*	Total number of trips observed in the mobile data during the day.	$V^* = 1,787,219$
v_t^*	Number of trips performed in the OD matrices at time step t.	$v_t^* \in \mathbb{N}$
$\hat{V}$	Number of trips performed by the population before rescaling during the day.	$\hat{V} = 3,085,661$
$\hat{v}_{it}$	Number of trips performed by individuals of category i at time step t.	$\hat{v}_{it} \in \mathbb{N}$
$\hat{s}_{ik}$	Indicator of category i describing joint socio-economic value k.	$\hat{s}_{ik} = n_i$ if category i has socio-economic description k, and 0 otherwise.
c	vector of rescaling coefficients.	$c \in \mathbb{R}^N$
A	Matrix representing the population, where each column is a category of agent.	Matrix of D rows and N columns
b	Vector of target counts for joint socio-economic values and number of trips by time step.	$b \in \mathbb{N}^D$

Table 4.9: Notations used in the formalization of the temporal calibration problem.

4.5 Results

In this section, we evaluate the realism of the synthetic population that we generated. Depending on the evaluation metric, we take as a reference the HTS, the census, or the mobile data. We also compare it to a **reference travel demand** generated using the work of Hörl and Balac (2021).

4.5.1 Rescaling coefficients

The rescaling problem has enough degrees of liberty so that Eq. 4.5 and Eq. 4.6 can be satisfied, if not exactly, then with a very small error margin. We compare the results of IPU with two versions of the optimization formulation: one with a low regularization coefficient $\beta = 5 \times 10^3$ that matches the target marginals almost perfectly but induces extreme values for the coefficients, and another with a high regularization $\beta = 5 \times 10^7$ that induces slight approximation in the counts of the less represented socio-economic categories, but presents a much more centered distribution of coefficients. In both cases the noise penalty is set to $\alpha = 5 \times 10^5$.

We first evaluate the approaches via the **Mean Square Error (E)** of the marginals obtained, which takes into account both the socioeconomic and the trip marginals:

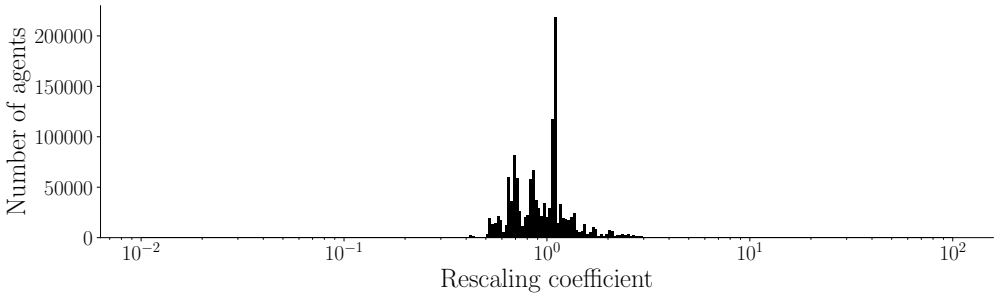
$$E = \frac{1}{T+K}\Big(\sum_{t=1}^{T}(\frac{1}{\hat{V}}\sum_{i=1}^{N}\hat{v}_{it}c_i - \frac{1}{V^*}v_t^*)^2 + \sum_{k=1}^{K}(\sum_{i=1}^{N}\hat{s}_{ik}c_i - \sum_{i=1}^{N}\hat{s}_{ik})^2\Big) \tag{4.15}$$

The other metrics we use for the evaluation aim to measure the presence of extreme values in the coefficients that would diminish the diversity of the population. We illustrate the distributions of the rescaling coefficients obtained for the various approaches in Fig. 4.11.

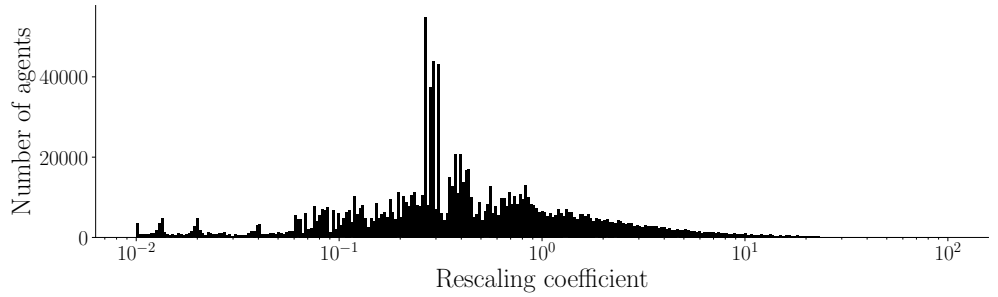
In particular, a problem that arises when not considering the regularization in the optimization problem is that the solution $(c_i)_{0<i\leq N}$ that is found contains mostly zeros, meaning that the population is more or less suppressed and replaced by a small subset of it. This is unacceptable as it destroys the diversity of the population that is precisely so hard to obtain. Coefficients that are below one are also at risk of being set to zero by the integerization step afterwards. The Truncate, Replicate, Sample approach for integerization can set a coefficient $c_i < 0$ to zero with probability equal to $1 - c_i$. We thus evaluate the approaches on the **expectation of number of deleted agents (S)**, given by:

$$S = \sum_{\substack{i=0 \\ c_i<1}}^{N} c_i \tag{4.16}$$

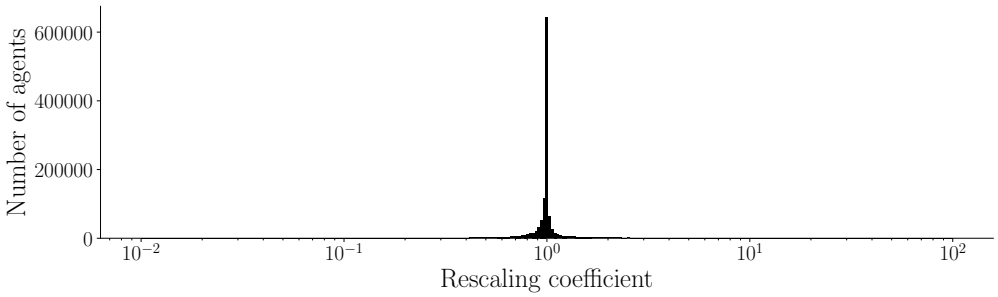
Finally, we consider two indicators to measure how the population as a whole is distorted by the coefficients. The first indicator is the **standard deviation (σ)** of the distribution of the coefficients, defined by:



(a) Histogram of the rescaling coefficients obtained by IPU.



(b) Optimization, $\beta = 5 \times 10^3$



(c) Optimization, $\beta = 5 \times 10^7$.

Figure 4.11: Histogram of the rescaling coefficients obtained by our various approaches

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(c_i - \bar{c})^2}, \tag{4.17}$$

where $\bar{c}$ is the mean of the coefficients. In practice, since the population was already at the right scale, the mean $\bar{c}$ is close to 1 so the standard deviation is equivalent to the root of the mean square error we get when comparing the rescaling coefficient to 1. This MSE is however not exactly what it considered as a regularization in Eq. 4.14, as the coefficients are grouped in the optimization problem. As a result, IPU gives a better, smaller σ than the optimization formulation.

Our second indicator for population distortion is the **Gini Coefficient (G)**. It corresponds to the area under the curve obtained when plotting the cumulative distribution of the

value of interest in the population, sorted in ascending order (Bendel et al., 1989). Such a curve is known as a Lorenz curve (Lorenz, 1905). The plot in our case is illustrated in Fig. 4.12. If we note ***AUC*** the area under the Lorenz curve, the Gini coefficient is given by:

$$G = 1 - 2 \times AUC. \tag{4.18}$$

In case of complete equality in the repartition, then the Lorenz curve is a diagonal and the Gini coefficient is $G = 0$. Absolute disparity is obtained when an infinite number of people have their value to 0 and one individual has everything. In that case, the Gini coefficient is $G = 1$.

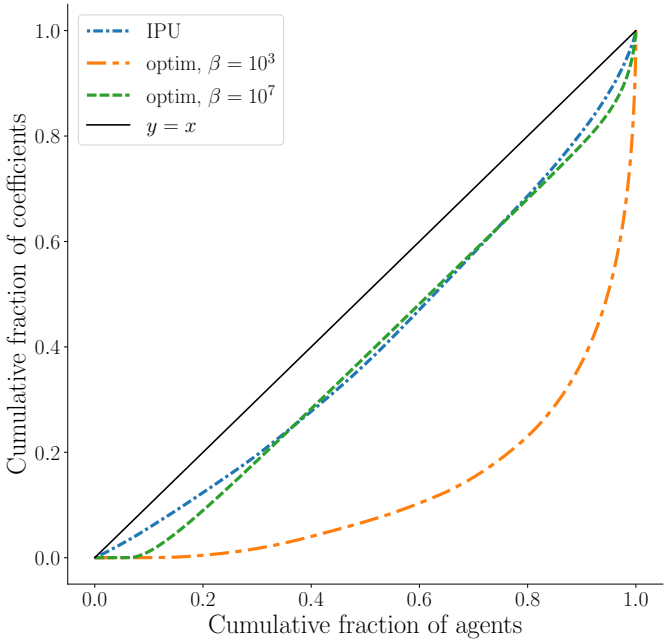


Figure 4.12: Lorenz curve of the rescaling coefficients obtained by our competing approaches. If we note AUC the area under the curve, then the Gini coefficient is given by $1 - 2 \times AUC$.

The results are summarized in Table 4.10. We see that despite having no rigorous justification, IPU is a very adequate approach to rescaling as it allows for both less suppression and less distortion of the population. By increasing the regularization hyper-parameter β in Eq. 4.14, we can obtain a distribution of coefficients similar to the one obtained *via* IPU but featuring discrepancies compared to the target marginals.

4.5.2 Socioeconomic composition of the population

In Fig. 4.13, we illustrate the fit to the target marginals for each of our solutions, and for the HTS for comparison. Each point is a combination of `gender`, `car availability`, `occupation`, and `age category`, with x-coordinates being the number of people with this joint value in the official census and y-coordinates the number of people with this joint values in the population sample. We observe a near perfect fit (all points are on the line $y = x$) for the populations rescaled by IPU and by the optimization problem with a low regularization. The small variations visible for low-volume marginals are due to the TRS integerization, which we

Metric	IPU	Optimization	
		$\beta = 5 \times 10^3$	$\beta = 5 \times 10^7$
$\boldsymbol{E}$ (Eq. 4.15)	5,312	**1,401**	20,057
$\boldsymbol{S}$ (Eq. 4.16)	12.95%	58.86%	**11.73%**
$\boldsymbol{\sigma}$ (Eq. 4.17)	**0.433**	3.514	0.734
$\boldsymbol{G}$ (Eq. 4.18)	**0.192**	0.759	0.214

Table 4.10: Evaluation of the various rescaling approaches for the temporal calibration.

can see is acceptably small. Higher regularization leads to an underfitting of the target, which corresponds to the higher value of the error E reported in Table 4.10. The HTS results, obtained by interpreting the micro-sample of the HTS (which is given with rescaling coefficients) like a synthetic population with activity chains, exhibits much larger discrepancies with the target marginals. The differences in socio-economic marginals is explained by the fact that as the HTS features few respondents, there is a high variance in the counts for a variable with a high number of values such as the joint socio-economic description. Note that the lower left points are missing, indicating that these categories are not represented at all in the HTS even though they exist in the census. Our synthetic population can then be seen as a version of the HTS that agrees with the census on the socioeconomic distribution of the population.

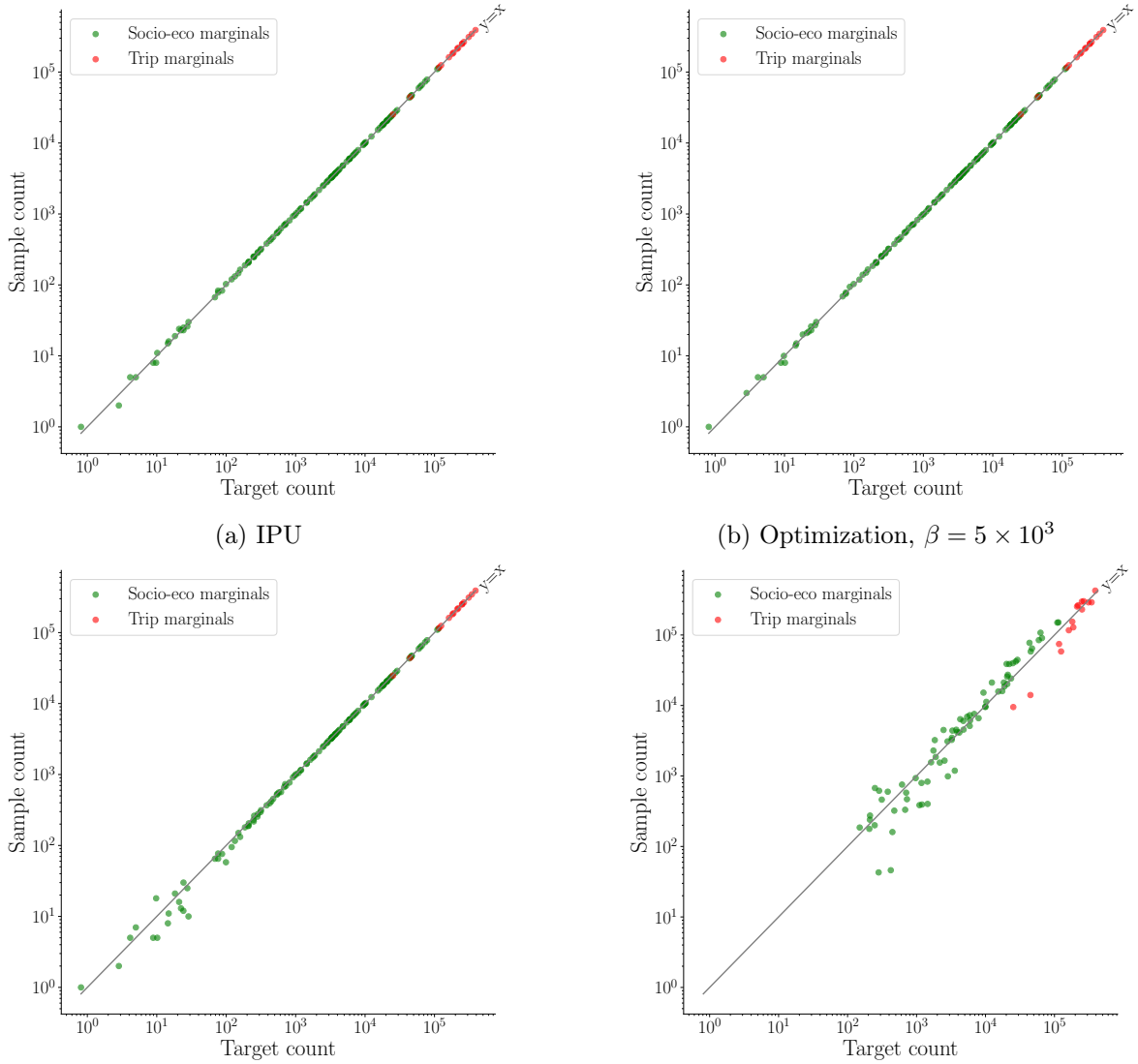
4.5.3 Number of trips by hour:

We consider that the true distribution of trips during the day is given by the mobile data and evaluate the synthetic demand in consequence. As discussed in Section 4.4, this is both an arbitrary and an optimistic stand, as we observe huge discrepancies between the mobile data and the surveys with little ways of verifying which is right. Still, we postulate that the reliability of mobile data is bound to improve in the future while surveys are likely to never reach more than a small fraction of the resident of an area, making it valuable to develop methods capable of using mobile data as reference.

The results are illustrated in Fig. 4.14, showing the fraction of trips taken at each time step as observed in the mobile data, our synthetic demand, the reference travel demand, and the HTS. We see that our approach accurately follows the distribution of the mobile data, while the reference travel demand is calibrated on the HTS, which features higher morning, noon, and evening peaks. As we did not performed the exact same pre-processing of HTS as the one used in the generation of the reference travel demand, we may report more discrepancies between the reference travel demand and the HTS compared to For better readability, we have redistributed the 15 time steps described in Table 4.6 into hours.

4.5.4 Distribution of agenda popularity:

Agendas are very unequal in popularity, as most people just go to work in the morning and back home in the evening. In this section, we consider that the agendas are characterized by the time step of each trip, their transport mode, and the purpose of their activities. In Fig. 4.15, we illustrate how the statistical matching and rescaling steps mostly preserve the



(a) IPU

(b) Optimization, $\beta = 5 \times 10^3$

(c) Optimization, $\beta = 5 \times 10^7$

(d) HTS

Figure 4.13: Matching of the socioeconomic marginals of the synthetic population (in red) compared to the HTS (in green).

popularity of the agendas compared to the HTS, which we consider to be the ground truth. We observe however a sensible under-representation of some of the least popular agendas to the benefit of others. Due to the low volume of the HTS, these agendas are actually held by only one survey respondent and appear with slightly distinct popularities only because of distinct scaling factors. Respondents with a popular socio-economic profile are then more frequently assigned in the statistical matching step than respondents with a more anecdotal profile.

We visualize the disparity in agenda popularity by plotting the cumulative distribution function of the agendas sorted by popularity. This results again in a Lorenz curve, which can be characterized by a Gini coefficient. Note that contrary to regular application of the Gini coefficient, it is not desirable to have a Gini coefficient of 0 in this case. It is rather more realistic to have a Gini coefficient as close as possible to the one of the HTS. We see in Fig. 4.16 that the shift in disparity is reasonably small.

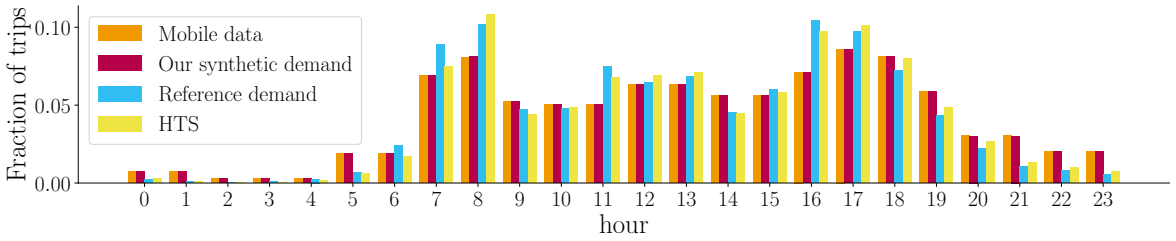


Figure 4.14: Distribution of trips taken during the day. For each hour, the first bar in orange is the fraction observed in the mobile data, the second in magenta is according to our synthetic demand, the third in blue is according to the reference travel demand, and the last in yellow is according to the HTS.

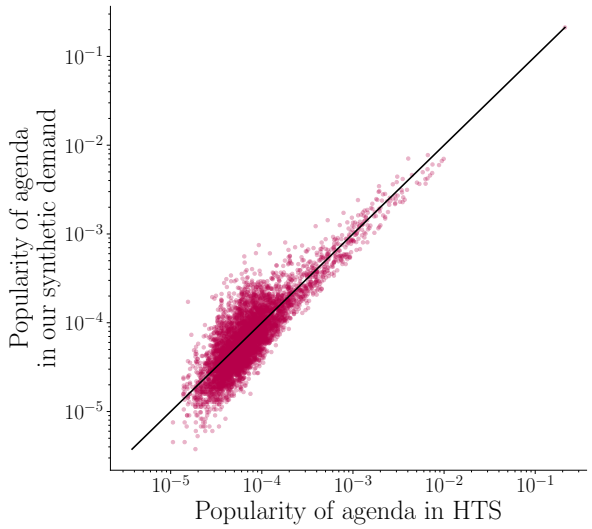


Figure 4.15: Popularity of agendas in our synthetic demand vs. popularity in HTS. The top-right dot is the empty agenda.

4.6 Discussion and conclusion

In this chapter, we reviewed the current state-of-the art for the first two steps of the population synthesis for travel demand, that are the generation of socio-economic factors and the assignment of activity chains. The approaches to these problems can have two main goals: either encourage diversity in the data by generating individuals and agendas that are likely but not explicitly observed in the input data, or calibrate the data with respect to some aggregate statistics.

Our contribution in this respect is the calibration of the temporal distribution of the synthetic demand to a distribution obtained from a tertiary source. By formalizing it as a hierarchical population problem, we can adjust the number of trips performed by the population during each time step of the day while retaining the socioeconomic composition of the synthetic population. This step can be applied *a posteriori* after assigning the activity chains, regardless of the exact approach used to this end. We perform the rescaling using IPU, which is a rather straightforward approach that lacks a formal justification, but we observe that it gives reliable, specifiable result with no tuning. In contrast, formalizing it as an optimization

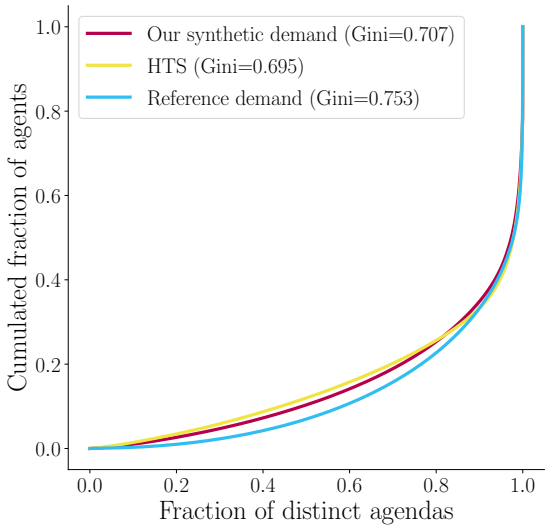


Figure 4.16: Cumulative distributions of the agendas from most to least popular, in our synthetic demand, the reference travel demand, and HTS.

problem requires to set a regularization coefficient that incurs a trade-off between a good fit of the marginals and a low suppression rate of the individuals.

Although it offers *a priori* no mathematical guarantee, IPU performs remarkably well for such calibration. The formulation of the problem as an optimization problem can be appealing by its tunability and more formal setting, yet give overall less satisfying results either in terms of fit to the target marginals or in terms of extreme values of the rescaling coefficients.

The realism of a synthetic population is determinant for its relevance in mobility studies. At the same time, this realism can only be measured in terms of coherence with separate data sources, where each source is often used as a reference for only one aspect. In the study case of this chapter for example, the mobile data is used as a reference for the temporal distribution of trips. Similarly, in the complete pipeline by Hörl and Balac (2021), the income database is separate from the micro-sample and a distinct step is necessary to assign incomes to the households such that the total distribution match the official results. In fact, even for socio-economic factors, for which micro-data is usually available, it is recommended to perform a deterministic rescaling such as IPF to ensure that the generated population, regardless of the model used to generate it, has aggregate statistics that match the official ones (Saadi et al., 2018). In that sense, the temporal calibration that we propose here is but one calibration step among others designed to take into account the various sources that we hold as ground truth in the population synthesis.

In this chapter, we disregarded the hierarchical aspect of the population on the basis that the household structures are seldom used in the assignment of activity chains, and so do not actually impact the resulting travel demand. This allows to shift the hierarchy for the purpose of the rescaling problem, where we consider the individual as the upper unit and the trips as the lower unit. In order to perform this calibration on a hierarchical population, further research is required to consider hierarchical populations with more than two layers of hierarchy.

Chapter 5

Spatializing synthetic travel demand with OD matrices

5.1 Introduction

In this chapter, we address the problem of assigning locations to the activities of the agents of a synthetic population. We follow-up on the methods described in Chapter 4, at the end of which we obtained a population of agents described by socio-economic variables and associated to activity chains. The activities are defined by their purposes, start time, and transport mode to get there, but have not yet been assigned locations.

Assigning these locations is still an open problem due to both the lack of available data on the matter and the very high dimensional nature of the problem: even in a simplified context of a study zone divided into a few hundreds of neighborhoods, a chain of locations can contain up to ten locations for a total of 10^{100} possibilities. Of course a strong structure exists in this distribution, and most of those 10^{100} possibilities would be physically impossible or at least highly unlikely. First of all, each individual can be expected to have a single location for all of their "home" activities, and in the general case the same is true for all their "work" and "study" activities. This leads the state-of-the-art to consider these **primary locations** separately from the others, as they are considered to be fixed anchor points independent from the rest of the agenda (Bowman and Ben-Akiva, 2001).

Popular methods to draw a secondary location between two primary activities consider the set of spatio-temporal points that can be reached between the end of the previous activity and the start of the next, given a certain maximum speed. This set is a **space-time prism** (Lenntorp, 1976) such as illustrated in Fig. 5.1. In travel demand synthesis, the space-time prism acts as a restriction of the set of possible locations to choose from (Yoon et al., 2012; Justen et al., 2013). Other, data-driven approaches such as Anda et al. (2021) model the joint distribution of locations, times, or transport modes as a probabilistic graph model, which parameters are then estimated with OD matrices and surveys.

In this chapter, we review the various approaches for the spatialization of synthetic travel demand. We consider a situation such as the state-of-the-art pipeline proposed by Hörl and

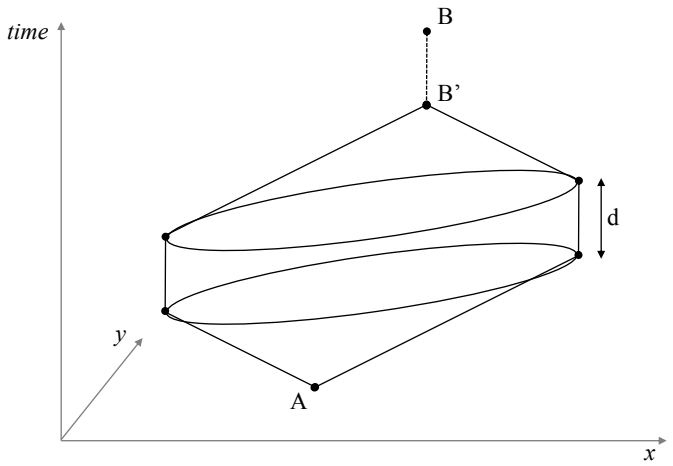


Figure 5.1: Illustration of a space-time prism between two anchor points $A = (x_A, y_A, t_A)$ and $B = (x_B, y_B, t_B)$, describing all the possible spatio-temporal points to start an activity of duration d. $B' = (x_B, y_B, t_B - d)$ is the most late point at which it is possible to start the activity without being late to meet anchor point B. Figure inspired from Lenntorp (1976).

Balac (2021), in which the spatialization in the last step of the travel demand synthesis. As such, the characteristics of the individuals and their activities are all considered fixed. We explore the use probabilistic graph models for the spatialization in such a context.

5.2 Spatialization background

5.2.1 Primary activity locations

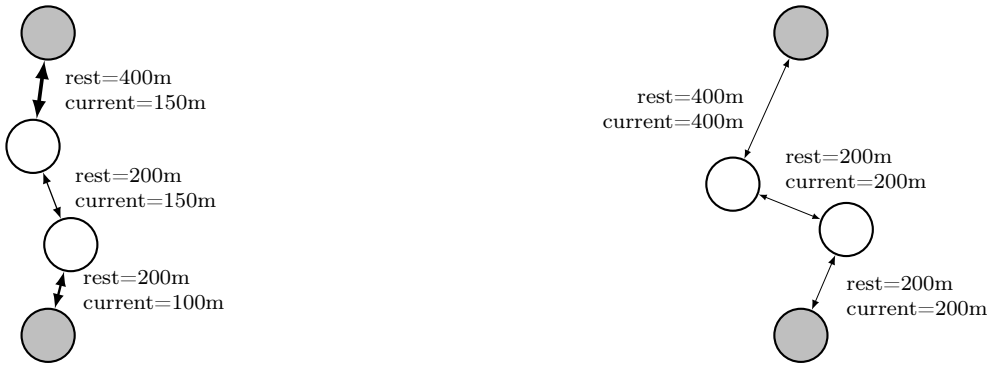
Activity locations are usually divided into primary locations (*e.g.*, home, work and education) and secondary locations for other types of activities (*e.g.*, shopping, leisure, personal). In home location research, the standard approaches rely on discrete choice models to replicate an individual's decisions in choosing a place to live (Wang et al., 2021; Hasibuan and Mulyani, 2022; Schirmer et al., 2014). When census data are available that specify home locations at the desired granularity, the straightforward approach is to use them directly (Hörl and Balac, 2021). Otherwise, the assignment of place of work and education is more appropriately handled by using a commute matrix when available (Hörl and Balac, 2021; Agriesti et al., 2022), or by relying on gravity models (Ahrens and Lyons, 2020), radiation models (Simini et al., 2012), or opportunity models (Liu and Yan, 2020). The primary locations can then be used as anchor points to assign secondary locations.

5.2.2 Secondary activity location

Choosing secondary locations proves to be much more difficult than primary locations when uniquely relying on HTS, as it does not cover enough people to be representative of precise locations. The state of the art in this area relies on discrete choice models (Huang and Levinson, 2015; Chen et al., 2018) restricted to space-time prisms (Justen et al., 2013), also known as activity spaces (Tsoleridis et al., 2023). Ma and Klein (2017) define a number of heuristic laws that agents would use in a discrete choice model, and use a Bayesian network to infer the actual heuristic used by each agent. However, these models are time-consuming

to tune and are difficult to generalize to other case studies.

Hörl and Axhausen (2021) propose a method for adjusting the distance between consecutive successive locations using a force model, simulating springs whose resting lengths are the distances between activities known from the HTS. The primary locations are anchor points, meaning that only the secondary locations move in order to satisfy all the trip lengths constraints. The equilibrium position serves as a basis for the activity locations. A discretization step then assigns the closest relevant infrastructure as the actual location chosen by the agent.



(a) Aligned, slightly disturbed locations

(b) Locations at equilibrium

Figure 5.2: Representation of the force model for the attribution of secondary locations. The grey locations are fixed. The direction of the arrows indicates the direction of the spring force, while the thickness indicates its strength.

A growing body of literature aims to generate trajectories from scratch, without consideration of the synthetic population (Kapp et al., 2023). Although they generally do not handle the semantic interpretation of the mobility, these approaches solve problem close to the localization of activity chains. Such models can be straightforward radiation model and opportunity models (Liu and Yan, 2020). Adequate random walks can also model the trajectory of individuals (Song et al., 2010), using a Lévy flight distribution for the trip distances and an **Exploration-Return** model based on the observation that people tend to return to the same place for the same kind of activity. Bindschaedler and Shokri (2016) proposes to represent trajectories in an abstract space and to use a Hidden Markov Chain as a decoder to generate semantically similar but distinct trajectories from any input trajectory. This also amounts to modeling trajectories by a probabilistic graph model. Anda et al. (2021) compare both approaches, and also propose the **Tour-explicit** model, which explicitly implements the fact that people return to the same place for a given activity. Recently, deep learning approaches have begun to show promising results for this task, such as Long-short-term-memory networks (LSTMs) (Lin et al., 2017), Generative Adversarial Networks (GANs) (Ouyang et al., 2018), or Variational Auto Encoders (VAEs) (Chiesa and Taraglio, 2022). Although the problem is close to the localization of activity chains, additional work is required to apply them to more extensive travel demand synthesis approaches. In particular, the drawing of the trajectories is not conditioned on a pre-defined activity chain or a socio-economic profile.

5.2.3 Mobile phone data for synthetic travel demand

The collection of mobile phone data is an ideal complement to surveys for generating synthetic travel demand (Bonnetain, 2022). Although not as detailed as surveys, such passive data provide more reliable location information and can be used to update older data to reflect rapidly changing mobility patterns. There has been a number of works using mobile phone data to synthesize a form of travel demand.

Using trajectories: Mobile phone trajectories can be directly interpreted as activity chains, interpreting each point as an activity (Zilske and Nagel, 2015). In this case, it is valuable to calibrate this *de facto* population with other data sources, such as link counts, which can be done by assigning scaling factors to the trajectories. Zilske and Nagel (2015) do this by creating several agents for each trace and feeding them to a simulator with the option to either follow the trace or do nothing. If the agents do nothing, they disappear from the mobility output. This allows for the simulator to perform the re-scaling of the population so that it matches observed link counts. Yin et al. (2017) use complete traces of mobile phones to train an input-output Hidden Markov Model able to generate activity purposes, along with parameters for models that will then sample their durations and locations: For example for secondary location, the model generates a distance from home and from work that will be used to sample the location. With the same goal of generating complete trajectories, Jiang et al. (2016) introduce the framework *TimeGeo* that find primary locations in mobile phone trajectories and the choice of secondary locations with a cascade method of analysis (Veneziano and Gonzalez, 2010). The same tasks can also be performed by an LSTM (Lin et al., 2017). Most generative approaches for the modelization of spatial trajectories, such as those described in Sec. 5.2.2, also consider mobile phone data as their working data. These approaches are useful to enrich mobile data trajectories, under the assumption that the spatio-temporal information they contain is reliable. It is worth noting that mobile phone data is not limited to spatio-temporal information. They also state the identifier of the receiver, which can be used to form a social graph for the synthetic population (Zhang et al., 2019). This approach is motivated by the idea that different social communities have different mobility behaviors. In some limited cases, mobile phone operators can also fetch the socio-demographic information associated to the mobile phone traces from their customer database. With this kind of enriched data, Ros and Albertos (2016) infer activity chains from CDR trajectories and assign them to a synthetic population derived from the Spanish census based on socio-demographic variables. It is also possible to train a model to predict basic socio-demographic features based on the CDR traces (Bwambale et al., 2019). All the approaches described in this paragraph use whole trajectories of mobile phone, which can be hard to come by due to understandable privacy issues. The enriched version of the traces associated to socio-demographic information is even more rarely available. In contrast, OD matrices are lighter, more manageable, and can be fully anonymized (Yin et al., 2015). This makes them safer regarding the privacy of the transportation users, and also more readily available as they can be safely disclosed by telecom operators.

Using OD matrices: There are multiple ways of making use of mobile phone OD matrices for the estimation of travel demand. They can first be used as reference data, for which other sources are rough estimates. For example, Huang et al. (2018) propose to transform the taxi

and subway flow counts of the city of Shenzhen, China, into total flow counts by choosing OD matrices from NSD as target. In this context, the explicative data is readily available on a regular operational basis but the NSD is only available for this specific calibration purpose. With the same goal of calibrating data to match the observation made from mobile phone, Bwambale et al. (2021) calibrate the trip production of the HTS of the city of Dhaka, Bengladesh, so that the number of trip generated for each zone corresponds to what is observed in the CDRs. These approaches are good examples of the general trend of using multiple data sources to generate more realistic and operationally useful mobility reports, and in particular with the assumption that the mobile data better represents the spatial distribution of a population than any other sources.

Aggregated OD matrices can be interpreted as conditional probability tables for probabilistic graph models. Anda et al. (2021) propose to generate spatialized agendas for the city of Singapore featuring the time of activities and the locations. The models used are the Exploration-Return model (Song et al., 2010) and a Tour-explicit model, in which a dependence structure is assumed among the variables of the agendas. Although these models can generate the time of trips and their structures along with the locations, they are not adapted to take into account other sources such as an HTS, and thus cannot generate transport modes and socioeconomic descriptions. Moreover, this approach benefits from using distinct models for various profiles of individuals. However, in the absence of socio-economical information, the individuals are defined only by their trajectory, ultimately making the approach reliant on a complete trajectory dataset. OD matrices can also be seen as forming a graph in which trajectories are paths. By decomposing the graph into a sum of path, it is possible to derive spatialized agendas (Ballis and Dimitriou, 2020). This approach is computationally prohibitive and could not be applied to the whole population of a city.

5.2.4 Evaluating synthetic mobility

Studies aiming at generating location chains from scratch usually evaluate the characteristics of the attractions of the areas, such as the proportion of visit (Bindschaedler and Shokri, 2016; Ouyang et al., 2018) or the radius of attraction (Schläpfer et al., 2021). A common evaluation of synthetic demand is based on the distribution of distances of the trips (Hörl and Balac, 2021), or even its origin-destination matrix when available (Fournier et al., 2021). Anda et al. (2021) evaluate their synthetic population by comparing eight marginal distributions obtained from the synthetic population with the ones obtained from the validation data. Among these eight marginals, four are directly connected to the localization of the output population. They are the population presence during the day, the total distance travelled by each individual, the sizes of their activity spaces, and their mobility entropy. The distributions of these criteria are compared using a version of the Earth Mover Distance (EMD) where the distance between two bins is always 1, which has the appeal of being a closed-form expression. However, closer inspection reveals that this closed form is simply equal to half of the L_1 distance. The choice of the metric to compare distributions in transportation is always hard, as the distributions are often sparse with a lot of possible values. Besides, the huge volumes when it comes to spatial data make the use of a general Earth Mover Distance computationnaly prohibitive. State of the art often refer to the L_1 distance (Gramaglia et al., 2017), or a variation of the L_2 distance such as the Mean Square Error (Farooq et al., 2013), the R^2 (Gong et al., 2023), or

Cramer's V (Sun et al., 2018). In this work, we preferably use the Kullback-Leibler divergence when applicable, which we compare to the Mean Square Error. In some cases, we also evaluate indicators of the distributions, such as the Gini coefficient, which we consider satisactory when they are realistic.

5.2.5 Positioning

Among the numerous approaches described above, very few allow to use mobile data jointly with an HTS. It is a missed opportunity to benefit both from exhaustive, aggregate location data from mobile phone and detailed individual descriptions from surveys. In this work, we propose to extend the state-of-the-art pipeline introduced by Hörl and Balac (2021) with a data-driven approach leveraging mobile data from the French telecommunications provider Orange. We use time-dependent OD matrices derived from mobile phone reconstructed trajectories (Bonnetain et al., 2021). We propose a spatialization step that replaces *ad-hoc* approaches by modeling agendas as a probabilistic graph model whose parameters are estimated from the OD matrices. This method is similar to using the models described by Anda et al. (2021), except that in our case the activity chains are already fixed. As such, our model does not aim at modeling the number of activities, the times at which they are performed, or when the individual is expected to go back home. By modelling only the location distribution, we can better evaluate the impact of the spatial data without having to handle the potential effects of other modelling details. Our work integrates the spatialization into the more general synthesis pipeline described by Hörl and Balac (2021). This pipeline can provide a synthetic travel demand data set describing individuals with socio-demographic attributes as well as structural activity chains. Our models must be used as the last step of such a pipeline, which is devoted to the spatialization. We also discuss a methodology to estimate these models, as their general, non-tree structures prevent us from directly interpreting the OD matrices as transition probabilities. Fig. 5.3 illustrates the positioning of our approach in the reference pipeline.

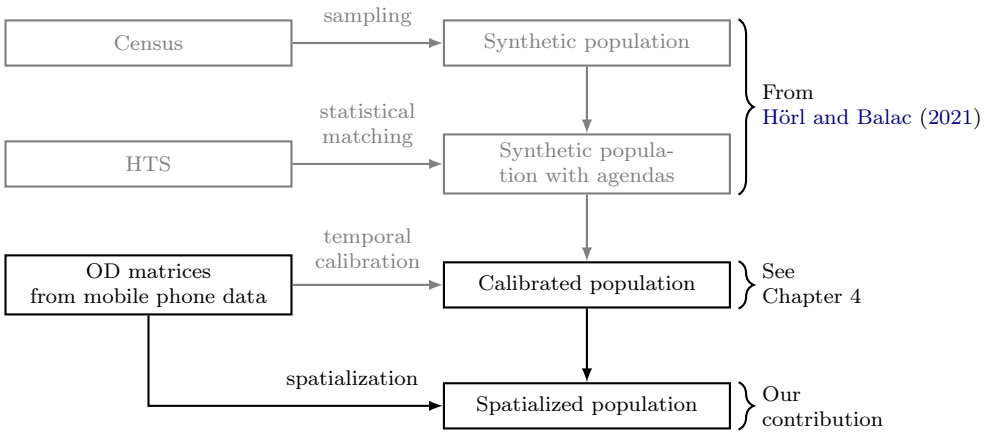


Figure 5.3: Positioning of the spatialization approach presented in this chapter.

5.3 Spatialization methodology

The purpose of this step is to assign locations to our agents so that they collectively form flows that resemble the structure observed in the OD matrices. The chains of locations for each individual should also be realistic with respect to their agendas: we assume that each agent has a unique location for all the activities with the same primary purpose. In particular, agents have only one home location, which is predefined in the synthetic population. We also assume that they have only one work location and one study location. This simplifying assumption does not take into account jobs that require traveling, such as delivery drivers, which represent 0.5% of the population according to the HTS and need to be filtered out in this study. More importantly, it does not handle working from home, which implies two possible locations for work activities and has become a widespread practice after the publication of our HTS.

5.3.1 Problem statement

We note Z_j^i the location of the jth activity of individual i. Z_j^i takes its values in the set $\mathcal{Z}$ of sub-zones defining a partitioning of the study area. Likewise, the purpose of the jth activity is noted P_j^i.

For each agent i, the constraint that the primary activities of the same purpose have the same location is noted `agenda constraint(i)`, which is satisfied if $Z_{j_1}^i = Z_{j_2}^i$ for all j_1, j_2 such that $P_{j_1}^i = P_{j_2}^i$ is a primary purpose. We also consider the constraint `spatial coherence(i)`, which is true if the chain of locations of agent i is coherent, that is Z_j^i is in the space-time prism defined by Z_{j-1}^i and Z_{j+1}^i. The flows described by the synthetic demand for each origin o, destination d, and time step t, are given by:

$$\hat{V}_{o,d,t} = \sum_{i=1}^{N} \sum_{j=1}^{\hat{v}_i} \mathbb{1}(Z_j^i = o, Z_{j+1}^i = d, T_j^i = t) \tag{5.1}$$

where $\mathbb{1}$ is the function that returns 1 if the condition is met, and T_j^i is the start time of the trip j which leads to activity $j+1$ (so T_j^i is also the ending time of activity j). $\hat{v}_i$ is the number of trips made by the individual i. Note that in these notations, only the Z_j^i are variables, while the P_j^i and T_j^i are determined by the agenda.

With these notations, the spatialization problem becomes the problem of finding Z_j^i that are realistic with respect to the agendas of the individuals and such that the resulting flows $\hat{V}_{o,d,t}$ resemble the ones observed from the OD matrices, noted $V^*_{o,d,t}$:

$$\begin{aligned} &\underset{\substack{Z_j^i, \\ 1 \leq i \leq N \\ 1 \leq j \leq \hat{v}_i}}{\operatorname{argmin}} \; L(\hat{V}_{o,d,t}, V^*_{o,d,t}) \\ &s.t.\ \forall i, 1 \leq i \leq N, \begin{cases} \texttt{agenda constraint(i)} \\ \texttt{spatial coherence(i)} \end{cases} \end{aligned} \tag{5.2}$$

In this problem, L denotes a metric of dissimilarity between the two matrices. Comparing

OD matrices is not trivial as they are sparse and have very unequal flows. We consider measures such as Kullback-Leibler divergence or R^2 to compare the estimated flows $\hat{V}_{o,d,t}$ with the flows observed from the OD matrices $V^*_{o,d,t}$. In each case, they correspond to a distance we aim to minimize between the observed OD matrix V^* and the OD matrix $\hat{V}$ formed collectively by our travel demand.

Although Problem 5.2 can be formulated as an optimization problem, its prohibitive number of variables calls for alternative approaches. The methods we propose in this chapter do not explicitely find the minimum of the dissimilarity $L(\hat{V}_{o,d,t}, V^*_{o,d,t})$, but instead use the counts $V^*_{o,d,t}$ as parameters for a sampling methods with the hope that the solution found has low dissimilarity with them. Using the mobile data in the sampling also ensures a form of the `spatial coherence` constraint, since the flows observed are by definition realisitic, so each location Z^i_j is both a realistic destination from the previous location Z^i_{j-1} and a realistic origin for Z^i_{j+1}. We argue that this helps ensure that each Z^i_j fits in a kind of fuzzy space-time prism between Z^i_{j-1} and Z^i_{j+1}.

Note that we rely on a single data source for both secondary and primary locations, such as work and study places. This is a stark contrast with state-of-the-art methods such as Yang et al. (2014) or Vitins et al. (2016) which tend to model the primary location well using commute matrices, but do not handle secondary locations. Our choice is justified by the fact that 45% of trips in the activity chains of our population are commute trips, whereas the mobile data contain the distribution of the sum of commute and non-commute trips. This would make it invalid to use it only for the latter purpose.

5.3.2 Spatialization

In this section, we describe the main contribution of the chapter, which is to assign a location to each activity of the synthetic population. We present successive structures of probabilistic graph model which differ in how they implement the agenda constraints described above. The first proposed approach does not respect these constraints, and serves as a benchmark for the best distance we can achieve.

Naive Forward Approach: We first implement a **naive forward** approach, where for each individual the locations Z_j are sampled from a Markov chain with transition matrices $M^j_{o,d} = p(Z_{j+1} = d | Z_j = o, T_j = t_j)$. Note that this Markov chain is inhomogeneous as the transition probabilities depend on the starting time T_j of the trip from activity j to activity $j+1$. An example is given in Fig. 5.4. These transition matrices are derived from the mobile phone OD matrices of the corresponding time step with each row normalised so as to be interpreted as a conditional probability law of the destination given the origin. For graphs that are trees, this estimation based on count of occurrences is the maximum likelihood estimator of the model.

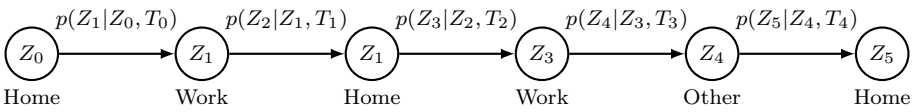


Figure 5.4: Naive assumption of the dependency of the locations on an example chain "h-w-h-w-o-h".

The naive forward spatialization consist simply in assigning the home location to Z_0, and successively drawing the next locations in order. It assumes that the first location Z_0 is home, which is the case for 98.3% of our agents. More importantly, agents are not guaranteed to go home when their agenda explicitly says so, which violates the agenda constraints. However, as it directly applies the transition matrices, the resulting travel demand will be as close as possible to the observed mobile phone data. The naive forward spatialization then serves as a baseline for the best fit we can get with a probabilistic graphical model.

Linked forward approach: A straightforward approach to implement the agenda constraint is simply to not draw the states that are fixed by the constraint (*i.e.,* a previous activity with the same primary purpose has already been assigned a location). We call this method as the **linked forward** approach. The corresponding graphical representation is shown in Fig. 5.5, where fixed states Z_2, Z_3 and Z_5 are determined by the agenda constraint instead of the transition matrix.



Figure 5.5: Dependency assumption of the linked forward approach.

We present this approach as it performs surprisingly well in our case, but we can expect it to generalize poorly on other data sets as it breaks the dependency chain: In our example, the location Z_4 depends on Z_3 but not on the fact that the agent should go home afterwards, potentially going far outside of the space time prism described by the home and work locations.

Linked MCMC Approach: The proper way of sampling a Markov chain such as Fig. 5.4 when some states are fixed would be to use a Markov chain Monte Carlo method. We perform Gibbs sampling in an approach called **linked MCMC**, where at each step we draw all the locations that share the same primary purpose, depending on their previous and next locations:

$$p(Z_j|Z_1, \ldots, Z_n, T_{j-1}) \propto \prod_{\substack{k, \\ P_j=P_k \text{ is primary}}} p(Z_k|Z_{k-1}, T_{k-1}) \times p(Z_{k+1}|Z_k, T_k)$$

This approach performs appallingly badly when the transition probabilities are estimated naively like in the forward approaches. In fact, the product of probabilities implies that the model is not a chain but rather a more general model where some states are linked, which we represent in Fig 5.6 as an undirected edge implying equality. As the model is not a tree anymore, the maximum likelihood estimator is not the occurrence counts of the transitions (Li, 2009).

We still include the linked MCMC approach in the benchmark as an illustration of the problems that arise when designing an arbitrary graph model. The next approach improves on this model by illustrating the estimation problem and proposing a heuristic solution.

Factor Graph Approach: The model presented in Fig. 5.6 describes a very general dependency structure, which can be formalized using a factor graph to better understand the

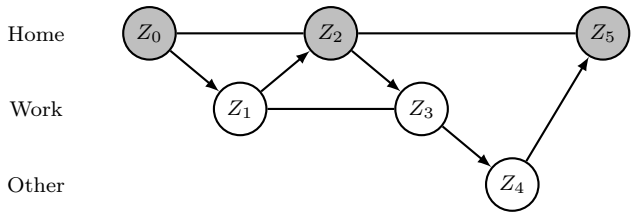


Figure 5.6: Locations distribution with explicit links between states sharing the same location for activity chain "h-w-h-w-o-h".

dependencies in the localisation problem (Jordan, 2015). A factor graph corresponds to a factorization of the joint distribution of the form:

$$p(Z_0, \ldots, Z_n) \propto \prod_i f_i(S_i),$$

where the factors f_i are each defined over a subset S_i of the variables $Z_0, \ldots, Z_n$. They can be interpreted as the generalization of conditional probabilities, as their product gives the joint distribution but only up to a scaling factor. The factor graph corresponding to Fig. 5.6 is given in Fig. 5.7, where a white square represents a factor defined over the variables it connects to.

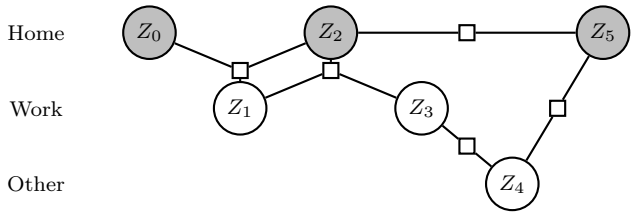
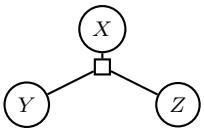
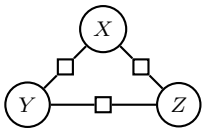


Figure 5.7: Factor graph representing the general factorization of Fig. 5.6.

We see in Fig. 5.7 that a round trip between two primary locations leads to factors of three variables: for example, the transitions "h-w-h" give a factor f_a linking Z_0, Z_1 and Z_2. Factors of three variables cannot be estimated from OD matrices, which only contain occurrence counts for pairs of locations. To obtain a graph that we can estimate from the OD matrices, we rely on the additional assumption that the factors of three variables can themselves be factorized in the form $f(X, Y, Z) = f'(X, Y) \times f''(X, Z) \times f'''(Y, Z)$, which corresponds to the graph transformation shown in Fig. 5.8.



(a) General form of a three factor $f(X, Y, Z)$.



(b) Sub-factorization as $f'(X, Y) \times f''(X, Z) \times f'''(Y, Z)$

Figure 5.8: Sub-factorization of a factor as three factors of two variables.

We obtain factors that either correspond to a transition from one location to the next, or factors between two linked states, in which case they are **equality factors** forcing the two variables to share the same value. The resulting factorization is described by Fig. 5.9, where white squares are factors corresponding to conditional probabilities, and black squares are

equality factors. Note that in this example, the factorization

$$f_a(Z_0, Z_1, Z_2) = f'_a(Z_0, Z_1) \times f''_a(Z_0, Z_2) \times f'''_a(Z_1, Z_2)$$

and

$$f_b(Z_1, Z_2, Z_3) = f'_b(Z_1, Z_2) \times f''_b(Z_1, Z_3) \times f'''_b(Z_2, Z_3)$$

result in two factors f'''_a and f'_b that depend on the same couple (Z_1, Z_2), as we see in Fig. 5.9.

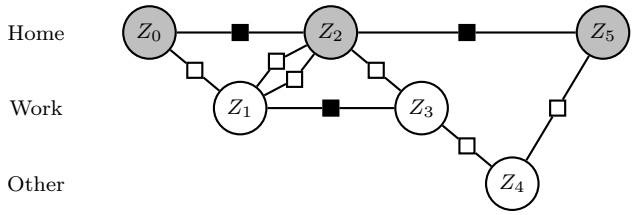


Figure 5.9: factorization of the location distribution with all assumptions. White squares denote factors that are conditional probabilities. Black squares are equality factors, forcing their neighboring states to be equal.

This model is the factor graph representation of the model implicitly used by the linked forward approach presented in the previous paragraph. It allows us to understand why the linked forward approach does not perform well and how to adapt it. As a general graph structure, the factors cannot be estimated by a simple counts of occurrences of each transition. A common estimator is the maximum pseudo-likelihood (Abbeel et al., 2006), which requires knowledge of the total joint distribution of the locations. Another approach based on piece-wise likelihood (Sutton and McCallum, 2005), could help estimate the parameters in the case that all factors are strictly positive, which is not our case as equality factors are a function returning either 0 or 1.

Instead, we base our estimation on the counts of occurrences with an heuristic adaptation based on the following observation: the transition law $p(Z_j|Z_{j-1}, T_{j-1})$ is a function equal to the product of all the factors between an activity linked to Z_{j-1} and an activity linked to Z_j. In the case of our example, there are 4 factors between the home and the work location. To make this product proportional to the actual volume of flows observed in the data, we can estimate each factor with the occurrence count and then apply a power $\frac{1}{n}$ to these n factors. We call approach as the **factor graph** approach. An intuition of the heuristic is given below, while a more detailed justification is given in Appendix 7.5. This approach implements the agenda constraints like the others, but also introduces a dependency on the next location when it is fixed by the structure of the HTS agenda. As such, we can expect that the flows described by the travel demand will be influenced by the HTS, and will not necessarily fit the input OD matrices perfectly. The different models are summarized in Table 5.1.

Intuition of the heuristic estimation of the factor graph approach: Let's consider a simple population of agents who all have the same commute agenda "h-w-h", as illustrated in Fig. 5.10. The marginal probability of two locations z_0 and z_1 is given by:

$$p(Z_0 = z_0, Z_1 = z_1|T_0) = \frac{1}{K} f_{0,1}(z_0, z_1) \sum_{z_2} f_{0,2}(z_0, z_2) f_{1,2}(z_1, z_2),$$

Model	Description
Naive forward	Sampling forward without returning home.
Linked forward	Sampling forward except when returning to a primary location.
Linked MCMC	Gibbs sampling from a chain with links between equal locations.
Factor graph	Gibbs sampling from a factor graph, factors estimated by an heuristic on the occurrences counts.

Table 5.1: Classification of the spatialization models used in this study.

where K is the normalisation factor. The equality factor $f_{0,2}(z_0, z_2)$ is 1 if $z_0 = z_2$ and 0 otherwise, so the sum simplifies to:

$$p(Z_0 = z_0, Z_1 = z_1 | T_0) = \frac{1}{K} f_{0,1}(z_0, z_1) f_{1,2}(z_1, z_0).$$

We want the marginal probability to be proportional to the volume observed in the data:

$$p(Z_0 = z_0, Z_1 = z_1 | T_0) = \frac{1}{K} V^*_{z_0, z_1, T_0}.$$

Note that in this example all the agents return home, so necessarily:

$$V^*_{z_0, z_1, T_0} = V^*_{z_1, z_2, T_1}.$$

So if we estimate the factors $f_{0,1}(z_0, z_1)$ by the volumes $V^*_{z_0, z_1, T_0}$, we obtain a an estimated probability law:

$$\hat{p}(Z_0 = z_0, Z_1 = z_1 | T_0) = \frac{1}{K} {V^*_{z_0, z_1, T_0}}^2,$$

which is proportional to the square of the volumes observed. The natural adaptation that arises is to consider the square root of the factors in that case. For a more general case where n factors can connect two sets of activities that each share a single location, see Appendix 7.5.

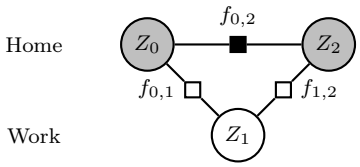


Figure 5.10: A simple commute agenda “h-w-h”.

5.4 Experiments and results

In this section, we apply our approach to a study zone covering Lyon and its surroundings as illustrated in Fig. 4.6. We first describe the chosen assessment metrics, and then use them to compare our synthetic travel demand to the one obtained with the state-of-the-art pipeline (Hörl and Balac, 2021) over the same region, which we referred to as the “reference demand”.

5.4.1 Assessment metrics

Evaluating synthetic demand is a problematic task due to the high dimensionality of the data and the lack of ground truth source. The only available data, in our case the HTS and the mobile phone OD matrices, are already used as input to the model. This leaves only internal validation, *i.e.,* comparing the synthetic population with the original input data (Anda et al., 2021; Fournier et al., 2021). Note that internal validation is not trivial as our input datasets are inconsistent with each other: it is then already a valuable result that the synthetic travel demand can match the mobile data on the temporal and spatial distributions while keeping agendas that are coherent with the HTS.

We expect the synthetic demand to match the mobile data regarding the distribution of trips during the day and the popularity of $o \to d$ flows, while the HTS serves as ground truth for the distribution of trip lengths and the popularity of activity chains. Note that although we proposed three distinct approaches, they differ only in spatialisation. For criteria that are not impacted by the actual activity locations, we refer to our approach as "our synthetic demand".

5.4.2 Distribution of distances

We compare the lengths of the trips taken in our synthetic travel demands to the distributions obtained from the HTS and the mobile data. The lengths are defined between the centroids of the origin and destination zones, meaning **static trips** staying in the same zone have a length of 0. We show the cumulative distributions of the trip lengths in Fig. 5.11. The forward methods, which apply in a very straightforward way the OD matrices, match the mobile data almost perfectly, while the factor graph approach allows the structure of the agendas to influence the probability of the destinations and gives a compromise between the mobile data and the HTS. The reference travel demand (Hörl and Balac, 2021) explicitly aims at respecting the distance distribution of the HTS, and succeeds. The gap for short distances can be explained by the differences in preprocessing of the HTS between this work and Hörl and Balac (2021). As expected, the linked MCMC approach gives unacceptable results, highlighting how the estimation of the transition probabilities can be non-trivial in graphical models. It features in particular around 60% of static trips, versus around 20% in the other sources. This can be explained by agenda patterns such as "h-z-h", where the linked MCMC approach assigns a location z with probability $p(z|h) \times p(h|z)$. This product, associated with the strong diagonal of our OD matrices, leads to a heavy overestimation of the probability that $z = h$.

5.4.3 Matching with the OD matrices

We compare the structure of the flows defined for each origin o, destination d and time step t observed from our synthetic demand $\hat{V}_{o,d,t}$ with the ones measured from the mobile data $V^*_{o,d,t}$. As the total volumes by time step $\hat{V}_t$ and V^*_t may not match since the temporal calibration only handles the distributions, we are interested in the normalized volumes: $\hat{y}_{o,d,t} = \frac{\hat{V}_{o,d,t}}{\hat{V}_t}$ and $y^*_{o,d,t} = \frac{V^*_{o,d,t}}{V^*_t}$. We compare the $\hat{y}_{o,d,t}$ with respect to $y^*_{o,d,t}$ in the scatter plots given in Fig. 5.12.

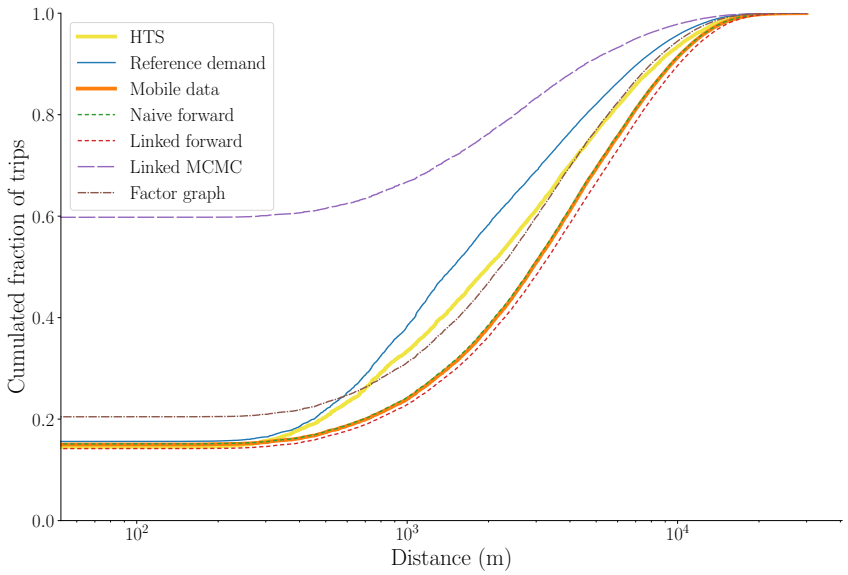


Figure 5.11: Distribution of trip lengths (in meters) between zones

We measure the fit to the diagonal in these plots with the R^2 metric:

$$R^2 = 1 - \frac{\sum_{o,d,t} (\hat{y}_{o,d,t} - y^*_{o,d,t})^2}{\sum_{o,d,t} (y^*_{o,d,t} - \overline{y^*})^2},$$

where $\overline{y^*}$ denotes the mean of all $y^*_{o,d,t}$. The R^2 is a direct measure of how well the synthetic travel demand performs compared to a demand where all flows had the same volume $\overline{y^*}$. Variations of it are a common measure to compare OD matrices (Yang et al., 2017), yet it tends to mostly measure the fit of the big flows, and associate only a small penalty to having small flows wrong. In our results, the 5% flows with the most error are systematically responsible for more than 90% of the total sum of square error, meaning the R^2 mostly measures the error of a few flows instead of a difference in the structure of the matrices.

In an attempt to better measure this difference, we interpret the $y^*_{o,d,t}$ and $\hat{y}_{o,d,t}$ as probability distribution $p(o, d|t)$. We compute the Kullback-Leibler (KL)-divergence between the synthetic and true distribution for each time step:

$$KL_t = \sum_{o,d} \hat{y}_{o,d,t} \log(\frac{\hat{y}_{o,d,t}}{y^*_{o,d,t}}).$$

Note that the KL divergence yields a finite value only if both distributions share the same support. We ensure that with a small additive smoothing $\epsilon = 10^{-30}$, which impacts the absolute values of the results but not the ratios between the approaches of the benchmark. We report the R^2 and the mean of KL_t across all time steps in Table 5.2.

The two measures interestingly do not give exactly the same conclusion, as they do not give the same importance to the same flows. In our case, the OD matrices feature a strong diagonal visible as the top-right cloud on the scatter plots of Fig. 5.12. The R^2 measures mostly those high volume flows, while the KL-divergence gives a more balanced measure of the fit

Model	R^2	KL
Naive forward	**0.76**	**9.39**
Linked forward	0.21	11.42
Linked MCMC	0.39	21.98
Factor graph	0.63	12.51
Reference	0.05	21.49

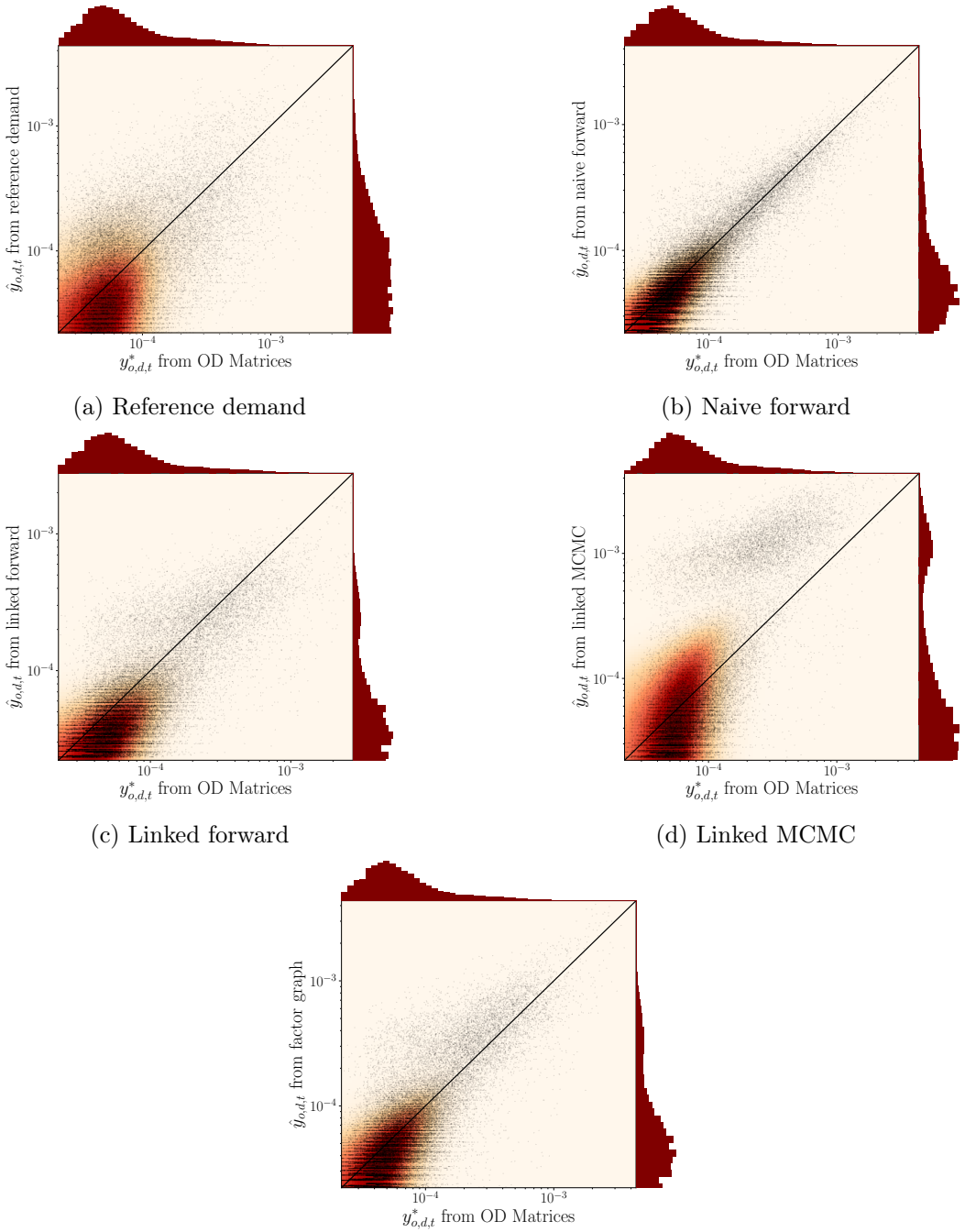
Table 5.2: Fit between the observed OD matrices and the synthetic travel demands of our benchmark.

between the distributions. In both cases, the naive forward method yields the best match to the mobile data, which is expected as it is the one that most straightforwardly applies the transition matrices as is. Its performance is also visible in the diagonal featured in Fig 5.12b. The linked forward and the linked MCMC approach introduce the agenda constraints and in consequence fit slightly less the input OD matrices regarding the KL divergence, which is visible in Fig 5.12c and Fig. 5.12e as a loose diagonal. Regarding R^2 however, the linked MCMC approach is much worse, presumably because it fails to estimate the largest flows correctly. Once again the linked MCMC approach performs outstandingly badly, even worse than the reference demand according to the KL divergence, even though the reference does not use the OD matrices in the first place. In particular, it heavily overestimates the biggest flows, which in our case happen to be the static flows. The reference demand, not benefiting from the mobile data, reflects how well synthetic data can match observed OD matrices using only *ad hoc* processes. This translates into a very loose distribution along the diagonal in Fig. 5.12a.

5.4.4 Matching with the commute matrices

Finally, we compare the approaches on their ability to approximate the official commute matrix of the area. This validation data is available at the commune level (equivalent to township), so we need to aggregate the travel demand. The commute matrix of a synthetic travel demand is obtained by extracting the home and work location of each individual, and aggregating it to the commune level. Note that the commute matrix is not defined for the naive forward approach, as the individuals do not have a unique home or work location. We illustrate in Fig. 5.13 the same kind of scatter plot where each point represents a couple (o, d) of two communes, and the x-coordinate and y-coordinate are the number of people living in o and working in d in the official survey and in the synthetic demand respectively.

As for the fit of OD matrices, we give the R^2 and KL-divergence of the distributions in Table 5.3. Unsurprisingly, the reference travel demand offers the best fit to the commute matrix with a R^2 of 0.99 and a KL-divergence of 0.23, as it explicitly uses it in its synthesis process. We notice in the scatter plot of the reference travel demand a consistent offset compared to the diagonal (Fig. 5.13a). Since it is in log scale, it indicates an underestimation of the commute volume by a constant fraction. This can be explained by the fact that the commute matrix does not represent trips, and it can be expected that a fraction of the population does not happen to go to work on the particular day of the agenda. The measures for the other



(a) Reference demand

(b) Naive forward

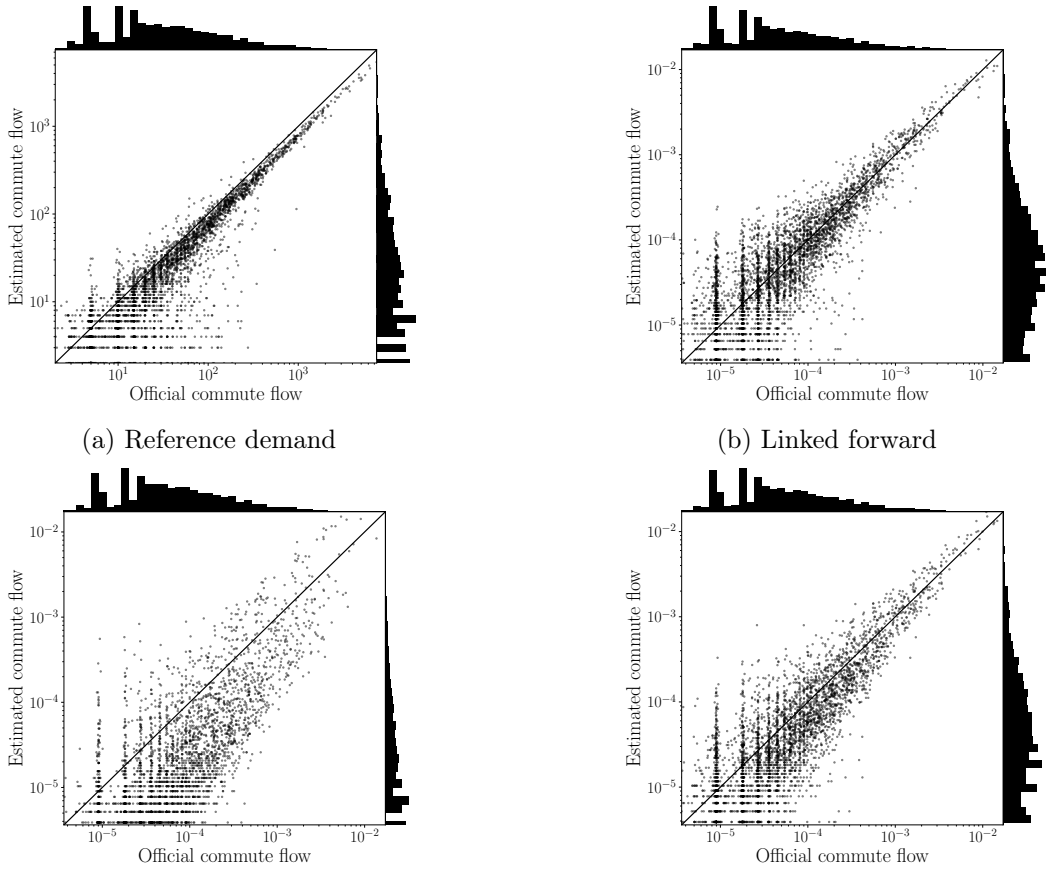
(c) Linked forward

(d) Linked MCMC

(e) Factor graph

Figure 5.12: Scatter plot of the OD matrices from synthetic demands compared to the mobile data.

approaches are rather undiscriminating, except for the R^2 confirming that the Linked MCMC is not a recommendable approach. Keep in mind however that out of the 8,836 possible pairs of communes, only 58 commute flows are responsible for 99% of the sum of square errors, and the KL divergence indicates that the structure in itself is not so decorrelated from the truth.



(a) Reference demand

(b) Linked forward

(c) Linked MCMC

(d) Factor graph

Figure 5.13: Scatter plot of the OD matrices from synthetic demands compared to the official commute matrix.

5.4.5 Discussion

Our approaches must be seen as a way to apply OD matrices when available, allowing for a more realistic travel demand provided the mobile data are of sufficient quality.

As noted above, our approach does not rely on predefined work and study locations because the mobile data that we use as input does not differentiate primary from secondary trips. Nevertheless, our approaches can still estimate the commute matrix to a reasonable level as we see in Fig. 5.13, considering that there is no specific calibration in this respect. Note that the OD matrices used for this study are issued from TRANSIT (Bonnetain et al., 2021), which includes a step to detect recurrent travels and enrich them with data from different days. This likely has a positive impact on the representation of "home → work" trips.

The most surprising result is probably how well the linked forward approach performs, although it breaks dependencies of the agenda. In particular, in the late evening when most agents are going home, the linked forward almost does not rely at all on the OD matrices and simply assigns the home destinations, to the risk of creating very unlikely trips. The factor graph approach, which is theoretically more correct, is not unequivocally better than the linked forward approach. This can be explained in several ways:

Model	R^2	KL
Naive forward	—	—
Linked forward	0.96	0.90
Linked MCMC	-0.33	0.78
Factor graph	0.92	0.73
Reference demand	**0.99**	**0.23**

Table 5.3: Fit between the official commute matrix and the synthetic travel demands of our benchmark.

- The current locations depend on more than the previous and next ones: It seems natural that people plan their day in advance and even have activities for which the location cannot be changed (such as visiting family), meaning it is not necessarily the location that changes but rather the inclusion of the activity in the agenda or not.
- The observations from mobile data and the agendas from HTS are incompatible, either because of the low quality of the data, or because the mobile data can by nature only observe stay points, as opposed to activities in the HTS which can be very short. In consequence, the trips to short activities are under-represented in the mobile data, and so are short trips due to their low spatial precision. As the factor graph approach allows the HTS structure to influence the transition probabilities, the incompatibility hampers its fit to the original OD matrices.

5.5 Conclusion on the spatialization

In this paper, we explored the use of OD matrices from mobile phone data as additional steps in the state-of-the-art pipeline for more realistic synthetic travel demand that mostly relies on an HTS. Mobile data are a valuable input for synthetic travel demand that aims both at being representative of a true population and based on recent observations. Full trajectory data are however a privacy liability and rarely available, so our approach focuses on using only OD matrices which can be fully anonymized while still providing valuable indicators of the travel demand.

As a first contribution of this chapter, we propose different approaches to use OD matrices for the spatialization of travel demand. This allows for a more realistic spatialization compared to state-of-the-art approaches that rely on heuristic methods. Each of these approaches corresponds to a trade-off between matching the HTS mobility statistics and matching the flows of the OD matrices, as these two sources present discrepancies that require arbitration.

A second contribution of our approach is then to provide a measure of those discrepancies, which are known to exist but hard to measure as the two sources describe different objects: the HTS describes individual agendas without representative localization of the population, while the OD matrices describe the localization of the population without individual coherence. By measuring the ability to compromise of our different approaches, we can give an estimate of how much these sources differ.

The limits of the spatialization approaches correspond to the limits of mobile phone data: A first shortcoming is that they do not differentiate between transport modes, which hampers the realism of the results as we know that different transport modes have very different mobility patterns. Another limitation is that they do not differentiate between commute and non-commute trips, which prevents us from using reliable commute matrices for the assignment of work and study locations.

A direct improvement on our method is then to provide transport mode-dependent, commute-dependent OD matrices. The detection of transport mode from mobile traces is a hard problem in urban environment, so additional approaches that derive this information from auxiliary data such as the HTS or the commute matrix may be necessary.

Our approach considers that each individual has only one work location, an assumption that is realistic in the 2015 population, which is the year of our HTS, but is becoming increasingly discutable with the emergence of working from home. The modelization of a possible choice to work from home or at the work place would also be a valuable addition allowing to take into account this new, important flexibility.

As our approaches are steps in the travel demand synthesis, they are also applicable in situations where an HTS is not available. In that case, the person agendas need to be generated using other approaches from the state-of-the-art. Various approaches have been proposed recently such as Bayesian networks (Joubert and De Waal, 2020; Sallard et al., 2021) or optimization-based scheduling choices (Pougala et al., 2023).

Chapter 6

Conclusion and research perspectives

6.1 Conclusion

In this section, we recapitulate the main contribution of the present work and we put them into the perspective of the research questions identified at the beginning of the thesis. The first main contribution of this work, presented in Chapter 3, explores the first question:

- **Question 1:** How can we guarantee the privacy in OD matrices while preserving their valuable information?

It is motivated by the need to make travel demand data more widely accessible: mobile data has until now been shared only confidentially, partly because it is proprietary, but mostly because it is personal. In addition, the impossibility to share the OD matrices is at odds with the increasing trend in research to rely only on publicly available data to ensure its reproductibility. In particular, recent travel demand synthesis frameworks such as Sallard et al. (2021) or Hörl and Balac (2021) aim at making travel demand synthesis accessible to anyone. Using mobile data in such pipelines requires a foolproof anonymization that could allow their communication on the same modalities than the other surveys on which the travel demand synthesis is based, *i.e.,* simply upon demand for research purpose.

Then, we explored the second question in Chapter 4 and Chapter 5:

- **Question 2:** How can we use OD matrices derived from mobile data to generate more realistic and comprehensive synthetic travel demand?

We based our approach on the state-of-the-art pipeline proposed by Hörl and Balac (2021), which handles each synthesis step from the population synthesis to the secondary locations in a modular manner, leaving space to develop other approaches for each specific problem. Such state-of-the-art pipelines rely greatly on HTSs, which are very detailed at the individual level and are generally considered reliable, but are still known to contain biases and cannot be sufficient for spatial analysis.

6.1.1 Anonymization of mobile data

A first step toward the public availability of OD matrices is k-anonymization. This criterion is widely accepted under the GDPR and by the French regulator CNIL as a satisfying condition for the anonymity of the data. There is a large number of solutions to generalize and suppress data in order to guarantee k-anonymity, and a variety of algorithms exist that restrain the search space at varying degrees before finding a solution. Most of these methods have been developed with the anonymization of health data in mind, which generally have numerous attributes but few possible values, and a low number of rows. In comparison, mobile phone data is characterized by its huge volume, even in the form of OD matrices, making impractical and expensive the use of computationally heavy algorithms. The anonymization of OD matrices requires a highly scalable approach both in terms of number of values and size of the data set. It can however take advantage of the fact that a time-specific OD matrix contains only two attributes.

We proposed two algorithms, ATG-Dual and ATG-Soft, that are adapted to this context to efficiently k-anonymize OD matrices. After decoupling the problem of the generalization of the origins and of the destinations, we solved each of the problems by formulating it as best pruning problem, looking for the pruning of a tree that maximizes the sum of the score of the nodes. This relatively easy problem becomes NP-hard when we introduce a constraint on the number of suppression we are willing to accept. Then, it can be formulated as a tree-knapsack problem, *i.e.,* a knapsack problem where the objects follow a dependency tree. To achieve high scalability, we solve the dual of the problem obtained by relaxing the suppression constraint, turning the problem into the maximization of its piece-wise linear Lagrangian function, which is solved efficiently by the Some Breakpoints Algorithm. This ATG-Dual approach can be further sped up by solving only the relaxed problem instead of the dual. The resulting ATG-Soft approach looses the guarantee on the suppression but introduces another property of interest, which is a hard constraint on the size of the individual generalizations. Our anonymization approaches are validated on a variety of data sets, illustrating both their superior scalability compared to solutions with a similar information loss, and their better information loss compared to solutions with a comparable computational cost.

Note that it could be argued that OD matrices do not constitute critical privacy threats, as individuals are only represented as collections of origin-destination trips that are impossible to link together. This already favorable context is in fact not sufficient in itself regarding both the European Union regulation and the privacy guarantees we could expect from a publicly available data set, but it is what allows to reasonably aim toward a completely foolproof anonymization that would give way to their public release.

6.1.2 Mobile data for synthetic travel demand

Mobile phone data is a promising source to measure the mobility of a whole population, which cannot be done by surveys. In Chapter 4, we considered that the temporal distribution of trips made during the day was better measured by the mobile data. Consequently, we proposed to use a hierarchical rescaling algorithm such as IPU to calibrate the temporal distribution of the synthetic population to this source. Methods of hierarchical population rescaling are usually used to assign scaling coefficients to households so that the socio-economic

composition of the people matches a given target. In our case, the upper unit is the individual, and the lower unit is the trip made by the individual. The hierarchical rescaling then allows us to match the target distribution of the mobile phone data while keeping the right socio-economic composition of the individuals. The resulting coefficients are evaluated on their capacity to match the target while suppressing as few agents as possible. Indeed, as the synthetic population is already of the right size, the rescaling coefficients are centered around one, making inevitable the assignment to some agents of coefficients below one, who then risk being suppressed from the data. As the rescaling problem is under-constrained, several solutions can be obtained that don't all imply the same suppression rate. We find that despite its lack of formal guarantees, IPU is superior to the optimization formulation in this regard.

In Chapter 5, we proposed an approach to spatialize the travel demand using OD matrices. In this regard, state-of-the-art approaches often handle primary locations with rather reliable sources such as commute matrices, but cannot handle the other, secondary locations that an agent can visit during the day. Those are usually decided based on a logic revolving around the primary locations as anchor points, and the secondary locations as auxiliary, flexible points that should disturb as little as possible the main agenda going from primary activities to primary activities. In our work, we interpret the OD matrices from mobile phone data as transition matrices between activities. These transition probabilities are used in a selection of models from the most straightforward that only uses the OD matrices, to the most developed that implements constraints from the agendas in its structure. This results in varying trade-offs between a travel demand that collectively fits the mobile phone data and one where the individuals are realistic compared to the surveys. The necessity to choose such a trade-off highlights the compatibility issues between the two sources. These issues are known to exist but have so far been hard to measure because of the different nature of the objects described: mobile data describes aggregated flows between static sessions of the customers of a telecom operator, while HTSs describe chains of activities separated by trips for a restricted number of survey respondents. By measuring the exact spatial fit of each of our possible trade-offs, we are capable of giving a quantitative measurement of these discrepancies.

Our approach is a direct step toward using OD matrices in the synthesis of travel demand. To a lesser extent, it is arguably a good step toward anonymization of the mobile data, as data that does not pertain to real individuals is less likely to be a danger to a particular person. This seemingly obvious statement actually deserves more nuance, as our population in this work is synthetic in the sense that it is rescaled and enriched with additional data, but is still largely based on existing census respondents. Even with fully synthetic agents, the data could still be subject to probabilistic attacks, which can be successful as long as the data is representative of the behavior of the target individual.

6.2 Perspectives

Our work opens a variety of research directions, which are worth investigating for the sake of making mobile data more available and better exploited in travel demand.

6.2.1 Better coupling origin-destination in anonymization

The anonymization of OD matrices that we propose in this work relies on the separation of the problem into the generalization of origins and the generalization of destinations. We highlighted the use of the hyper-parameter `v_target`, which balances the coarseness of origins and destinations, with the goal of finding anonymization solutions that don't favor one or the other. It is likely that a best value for `v_target` exists, and that it depends on the structure and the number of flows of the matrix. The current solution, choosing a `v_target` for a whole set of OD matrices by evaluating a small subset, is unsatisfactory. An obvious alternative would be to solve directly for the joint problem by computing, for each node of the generalization tree of the origin, the generalization of destinations. This is currently prohibitive in terms of computational cost but additional implementation efforts might result in a joint-problem algorithm that runs in reasonable time. In particular, the implementation has been made in python in a logic of benchmarking multiple solutions, but an operational usage would require an implementation in a faster language. The tree structure as well as the solving of separate problem for each aggregated origin make the algorithm suitable for efficient parallelization, which would also be of great interest in the solving of the joint generalization problem.

6.2.2 Hierarchical population

Our synthetic travel demand has explicitly dropped the household structure of the population. It is well documented that people of a same household have complementary agendas: we could expect that if one parent drives the children to school on a given day, then the other does not need to do it, or that someone doing the groceries implies that the other persons in the household will not have to do it. These considerations affect the purposes of the activities, but also the time and the locations, since an activity such as "fetching the children at school" has the same location as the "school" activity of the children, and a time depending on the hour at which school finishes. The agendas stated in our HTS detail forty-two distinct purposes which are mapped into the four simplified category that we use in this work. This detailing of the purposes allows in particular to detect the interacting situations such as accompanying someone else to do an activity. The research in synthetic population is also well documented on the hierarchical population problem, where the goal is to create a population of households that match a target distribution, while simultaneously accounting for the target distribution of the individuals. Our spatialization models are already designed to take into account location interactions *via* the introduction of links. Given all the agendas of the individuals of a household, a straightforward adaptation of our approaches would be to represent all the agendas of the household as a single graph where the locations that we know equal across agendas (such as the location of the school where the children go and where the parents fetch them) are linked in the same way two work locations are linked in a single agenda. As the graphs would be larger, the warm-up phase of the Markov Chain Monte Carlo sampling would be longer than for drawing independent agendas, but the computing time would likely remain in the acceptable range of one hour for the 1.3M agendas.

However, an important missing piece before fully taking into account household interactions is a household-aware assignment of agendas. The state of the art in agenda generation mostly relies on either mobile phone trajectories, such as Anda et al. (2021), or HTSs, such as Sallard et al. (2021); Pougala et al. (2023). In both cases, the focus is on reconstructing

the distribution of the individual's chains, without considering how the chains interact with other individuals in the same household. Very little work has been done on the generation of interacting activity chains (Anggraini, 2009; Vo et al., 2020, 2021). Note that household-aware agenda assignment would also require the HTS to describe in much more detail the purposes of the activities, as it would become necessary to know if an activity corresponds to meeting another member of the household. Statistical matching has no obvious possible generalization to make it into a hierarchical version, and the necessary further detailing of the purposes would risk creating situations where most agendas are held by only one survey respondent. In our work, we see that the popularity of such agendas is not respected by the statistical matching, so we risk generating a population of individuals whose agendas are nothing like the agendas of the true population. In addition, the temporal calibration we perform in our work already makes use of a hierarchical rescaling where the upper entity is the individual and the lower entity is the trip. The addition of a third hierarchical layer for the households can be handled by adapting the already existing approaches for two-layer hierarchical populations, but this problem has not been fully explored yet.

6.2.3 Better estimation of factor graphs

The factor graphs used for the spatialization of synthetic travel demand have a very general structure that make invalid the straightforward interpretation of OD matrices as transition probabilities. Instead, we relied on an illustration of the problem to propose an adapted parametrization, applying an exponent based on the number of transitions between two identical sets of locations. We illustrated how this adaptation helps improving the performances of the spatialization in terms of fit to the original OD matrices, but this simple estimation problem could be more formally solved using advanced parameter estimation approaches, such as piece-wise likelihood (Sutton and McCallum, 2005), which are specifically designed to estimate graph models from data where only pieces of the joint distribution are observed. Those methods are however valid only for factor graphs where all factors are strictly positive, which is not the case in our situation because of the constraint factors. Another way of handling the problem could be to associate each node of the graph to a location rather than to an activity. This approach would avoid the problems caused by the equality links. However, each factors would then need to be estimated from several OD matrices instead of only one with our current models, for example, if several "home $\rightarrow$ work" transitions happen in the agenda. The model would give no indication on the correct way to estimate those factors from such a collection of OD matrices.

6.2.4 Better matching emission maps from mobile data

The spatialization method presented in Chapter 5 uses the mobile data as transition probability with the justification that it transforms "popular" observed trips into "probable" trips to draw from. However, this approach does not guarantee to fully reproduce the patterns observed in the mobile data. In particular, the emission maps of the synthetic travel demand for each time step has no guarantee to correspond to the emission maps of the mobile data. This is due to the fact that the agents are sent to locations independently of the time of their next trip, although it impacts the synthetic emission map of the other time steps. As a result, if we consider a zone z and a time step t_0, for which the mobile data state that n agents leave

the zone, then there is actually no guarantee that n agents whose agenda contain a trip at time step t_0 are present in zone z. The current temporal calibration presented in Chapter 5 only guarantees that the population as a whole makes the right number of trips, but it does not consider where these trips come from.

One way to ensure that the agents who make a trip at a given time step t_0 are spatially distributed where there are needed would be to use custom transition matrices that depend on the time step of the next trip of the agents. For example, an agent who leaves home at 8am and whose next trip is at 5pm should preferably be sent in a place that emits a lot of trips at 5pm. Conversely, an agent who also leaves home at 8am but whose next trip is at 9am could be needed elsewhere from our point of view.

More formally, this would amount to adding an explanatory variable to the transition matrices. Instead of being of the form $M^*(o, d, t)$, with a probability depending only on the origin, the destination, and the current time step, they would be of the form $\hat{M}(o, d, t_1, t_2)$, depending on the current time step and the next one planned in the agenda.

Adding this extra variable can be done via a method similar to IPF, in the sense that it implies finding a joint distribution under the constraint of some fixed marginal distributions. In this case, the first marginals respected by the enriched matrices would be that they should sum to the initial OD matrices:

$$\forall t_1, \sum_{\substack{t_2 \\ t_2 \geq t_1}} \hat{M}(o, d, t_1, t_2) = M^*(o, d, t_1) \tag{6.1}$$

The other constraint that these enriched matrices would have to respect is sending the agents where they are needed depending on the trip time planned on their agendas. This means that the destination maps of the matrices $\hat{M}$ sum up to the emission maps of the next OD matrices. More formally, we note the destination maps $\hat{M}_d$, which are obtained by marginalization over the origins:

$$\hat{M}_d(d, t_1, t_2) = \sum_o \hat{M}(o, d, t_1, t_2) \tag{6.2}$$

Likewise, the emission maps M^*_o of the mobile phone OD matrices are obtained by marginalization over the destinations:

$$M^*_o(o, t_1) = \sum_d M^*(o, d, t_1) \tag{6.3}$$

The synthetic agents who make a trip at time step t_1 start from a zone z either because their last trip sent them in z, or because this is their first trip of the day. We note $H(t_1)$ the map counting the agents whose first trip is at t_1. Then, respecting the emission maps observed in the mobile data means:

$$M^*_o(t_2) = \sum_{\substack{t_1 \\ t_1 \leq t_2}} \hat{M}_d(t_1, t_1) + H(t_2). \tag{6.4}$$

Equations 6.1 and 6.4 describe constraints the marginal distributions of the enriched transition matrices $\hat{M}(o, d, t_1, t_2)$. These constraints can themselves admit multiple solutions, and there are multiple possible OD matrices that respect a given destination map. Solving this

problem with rescaling solutions such as IPF or generalized raking could give an adequate estimation for $\hat{M}(o, d, t_1, t_2)$. This would eventually allow us to assign locations to each of our agents depending on where we logistically need them to guarantee the right volumes of trips in every regions of the study zone. However, note that adding this step only aims at respecting an aggregate indicator given by the mobile data, but it does not ensure a better individual coherence of the location chains. As such, care should be taken as it risks encouraging unrealistic individual behaviors if they are necessary in order to match the constraints that have been set at the population level.

6.2.5 Better taking into account commute matrices for primary locations

The strength and weakness of OD matrices derived from mobile phone data is that it captures the movements of a large fraction of the population, regardless of their transport mode, socio-economic profile, or trip purpose. In particular, 31.8% of the trips in our HTS are commute trips, which are known to be very different in their characteristics than non-commute trips. They are made only by the specific part of the population that is either studying or working, they happen at specific times of the day, and they are usually longer (Fig. 6.1): the median distance of commute trips in our HTS is 2,450 meters, versus 1,140 meters for non-commute trips. Our spatialization approaches indifferently draws locations for work and

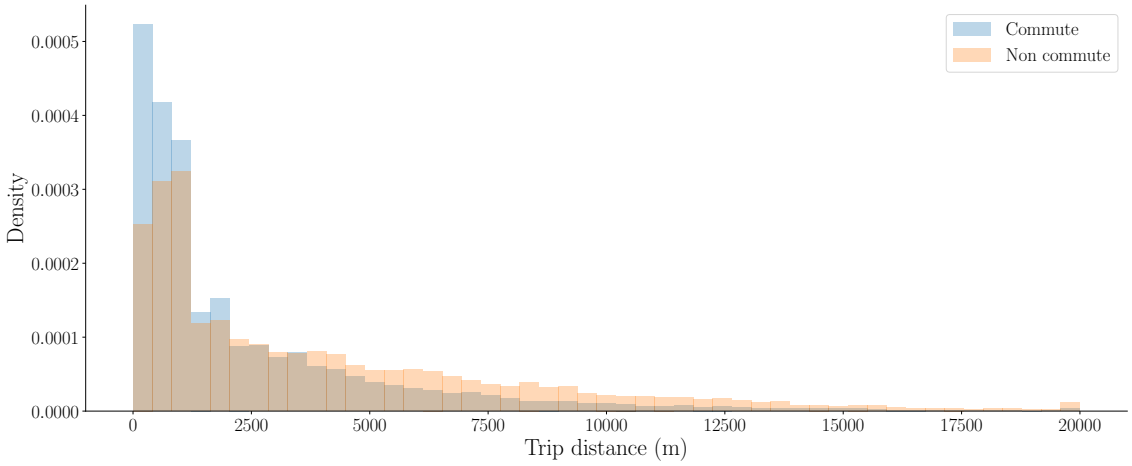


Figure 6.1: Distribution of the distance of commute trips and non-commute trips in our HTS.

non-work trips with the same transition matrices, based on the remark that mobile phone data cannot differentiate between the two and so the OD matrices contains both. Despite this, we have shown that the commute patterns of our synthetic travel demand are reasonably realistic. However, additional steps could be taken to ensure a perfect fit between the synthetic travel demand and a reference commute matrix.

A first solution could be to add a spatially-dependent rescaling step after the spatialization. This step would resemble the one described in Chapter 4, expect that the assigned work location would be an additional feature on the same level than socio-economic features. Such a step would ensure the calibration of the commuting patterns while maintaining fixed all the other variables we include in the rescaling, but at the risk of distorting all the other aspects of the population that are not explicitly considered in the rescaling.

Another solution would be draw initially the primary locations based on the commute

matrix, and use the mobile phone data in the spatialization step only for the secondary locations. This would likely degrade the fit between the synthetic OD flows and the observed mobile phone OD matrices. The realism of the resulting trajectories could also be impacted, as it would mean drawing secondary locations from transition matrices that represent trips to both primary and secondary locations. This solution would be adequate if the OD matrices we have at our disposal were separated between "commute" and "non-commute". In fact, a variety of conditioning variables could be considered in the OD matrices in order to improve on the realism of the travel demand, as discussed in the next perspective.

6.2.6 Additional conditioning on the spatialization drawing

The spatialization drawing is dependent on the previous and next locations, meaning that the location that is sampled for each activity of each individual is a likely destination for a trip leaving from the previous activity, and a likely origin for a trip going to the next one. Since the probabilities are given by the OD matrices, a likely trip in this context is a trip that numerous people have been observed to make. The probability law of a location given the previous and the next then corresponds to the actual movement possibilities inside the city, drawing a kind of fuzzy space-time prism. However, the drawing does not take into account other characteristics of the trips that are available from the HTS agendas, such as the distance, the transport mode, or the purpose of the destination activity. The two latter are common features of OD matrices, although they were not available in the OD matrices that we used in this work. However, it is possible to condition the drawing of the destinations on the distance using rejection sampling, which allows to draw the destination from the OD matrices among only those that are at the desired distance. This would ensure that the distribution of actual trip lengths in the synthetic travel demand matches the distribution of trip lengths stated in the agendas of the synthetic population. In order to ensure the calibration of the synthetic travel demand to the HTS, it would also be necessary to make sure that the agendas of the synthetic population present the same length distribution as the HTS. This can be done by adding a variable in the description of the trip at the temporal calibration level described in Chapter 4.

6.2.7 Better conciliation of mobile data and HTS

In order to use the mobile data in the demand synthesis, we relied on a preprocessing that illustrated how trips such as small errands on foot are under-represented by mobile data. In a more general context, in order to use the data in the best conditions, it would be valuable to characterize more precisely which kind of trips are well represented by mobile data, and simply rely on other sources for the rest. This would help alleviate the problem of definitions of "HTS trips" versus "Mobile phone trip". Another step of the HTS preprocessing assumes that the respondents have only one work place, an assumption that is relevant for a pre-Covid study, but is becoming more and more discutable with the advent of work from home. Futur travel demand synthesis should not rely on this assumption and integrate the possible choice of working from home in the spatialization. In a more general context, each source of spatial information for transportation applications is characterized by specific collection conditions which influence the exact type of trips or activities that is measured and best represented. A valuable direction of research would be to make explicit these characterizations in order to

apply each source to the context where it is best fitted.

Bibliography

Pieter Abbeel, Daphne Koller, and Andrew Y. Ng. Learning factor graphs in polynomial time and sample complexity. *J. Mach. Learn. Res.*, 7:1743–1788, dec 2006. ISSN 1532-4435.

Osman Abul, Francesco Bonchi, and Mirco Nanni. Never walk alone: Uncertainty for anonymity in moving objects databases. In *2008 IEEE 24th International Conference on Data Engineering*, pages 376–385, 2008. doi: 10.1109/ICDE.2008.4497446.

Osman Abul, Francesco Bonchi, and Mirco Nanni. Anonymization of moving objects databases by clustering and perturbation. *Information Systems*, 35(8):884–910, 2010. ISSN 0306-4379. doi: https://doi.org/10.1016/j.is.2010.05.003. URL https://www.sciencedirect.com/science/article/pii/S0306437910000426.

Muhammad Adnan, Francisco C Pereira, Carlos Miguel Lima Azevedo, Kakali Basak, Milan Lovric, Sebastián Raveau, Yi Zhu, Joseph Ferreira, Christopher Zegras, and Moshe Ben-Akiva. Simmobility: A multi-scale integrated agent-based simulation platform. In *95th Annual Meeting of the Transportation Research Board Forthcoming in Transportation Research Record*, volume 2. The National Academies of Sciences, Engineering, and Medicine Washington, DC, 2016.

Zack Aemmer and Don MacKenzie. Generative population synthesis for joint household and individual characteristics. *Computers, Environment and Urban Systems*, 96:101852, 2022. ISSN 0198-9715. doi: https://doi.org/10.1016/j.compenvurbsys.2022.101852. URL https://www.sciencedirect.com/science/article/pii/S0198971522000965.

Serio Agriesti, Claudio Roncoli, and Bat-hen Nahmias-Biran. Assignment of a synthetic population for activity-based modeling employing publicly available data. *ISPRS International Journal of Geo-Information*, 11(2), 2022. ISSN 2220-9964. doi: 10.3390/ijgi11020148. URL https://www.mdpi.com/2220-9964/11/2/148.

Ahmed Ahmouda, Hartwig H. Hochmair, and Sreten Cvetojevic. Using twitter to analyze the effect of hurricanes on human mobility patterns. *Urban Science*, 3(3), 2019. ISSN 2413-8851. doi: 10.3390/urbansci3030087. URL https://www.mdpi.com/2413-8851/3/3/87.

Achim Ahrens and Sean Lyons. Do rising rents lead to longer commutes? a gravity model of commuting flows in ireland. *Urban Studies*, 58:004209802091069, 05 2020. doi: 10.1177/0042098020910698.

Laura Alessandretti, Luis Guillermo Natera Orozco, Meead Saberi, Michael Szell, and Federico Battiston. Multimodal urban mobility and multilayer transport networks. *Environment and Planning B: Urban Analytics and City Science*, 50(8):2038–2070, 2023. doi: 10.1177/23998083221108190. URL https://doi.org/10.1177/23998083221108190.

Alberto Aleta, Sandro Meloni, and Yamir Moreno. A multilayer perspective for the analysis of urban transportation systems. *Scientific Reports*, 7(1):44359, Mar 2017. ISSN 2045-2322. doi: 10.1038/srep44359. URL https://doi.org/10.1038/srep44359.

Christian Alis, Erika Fille Legara, and Christopher Monterola. Generalized radiation model for human migration. *Scientific Reports*, 11(1):22707, 2021.

Cuauhtemoc Anda, Sergio A. Ordonez Medina, and Kay W. Axhausen. Synthesising digital twin travellers: Individual travel demand from aggregated mobile phone data. *Transportation Research Part C: Emerging*

Technologies, 128:103118, 2021. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2021.103118. URL https://www.sciencedirect.com/science/article/pii/S0968090X21001376.

R. Anggraini. *Household activity-travel behavior : implementation of within-household interactions*. Phd thesis 1 (research tu/e / graduation tu/e), Built Environment, 2009. Proefschrift.

Theo Arentze, Frank Hofman, Henk Mourik, and H.J.P. Timmermans. Albatross: Multiagent, rule-based model of activity pattern decisions. *Transportation Research Record*, 1706:136–144, 01 2000. doi: 10.3141/1706-16.

Fereshteh Asgari, Alexis Sultan, Haoyi Xiong, Vincent Gauthier, and Mounim El Yacoubi. Ct-mapper: Mapping sparse multimodal cellular trajectories using a multilayer transportation network. *Computer Communications*, 95, 04 2016.

David Ashley, Tony Richardson, and Dave Young. Recent information on the under-reporting of trips in household travel surveys. In *Australasian Transport Research Forum (ATRF), 32nd, 2009, Auckland, New Zealand*, 2009. URL https://api.semanticscholar.org/CorpusID:96518168.

Joshua Auld, Michael Hope, Hubert Ley, Vadim Sokolov, Bo Xu, and Kuilin Zhang. Polaris: Agent-based modeling framework development and implementation for integrated travel demand and network and operations simulations. *Transportation Research Part C: Emerging Technologies*, 64, 08 2015.

Kay W. Axhausen and Raimund Herz. Simulating activity chains: German approach. *Journal of Transportation Engineering*, 115(3):316 – 325, 1989. doi: 10.1061/(ASCE)0733-947X(1989)115:3(316). URL https://doi.org/10.1061/(ASCE)0733-947X(1989)115:3(316).

Danya Bachir, Ghazaleh Khodabandelou, Vincent Gauthier, Mounim El Yacoubi, and Eric Vachon. Combining bayesian inference and clustering for transport mode detection from sparse and noisy geolocation data. In Ulf Brefeld, Edward Curry, Elizabeth Daly, Brian MacNamee, Alice Marascu, Fabio Pinelli, Michele Berlingerio, and Neil Hurley, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 569–584, Cham, 2019. Springer International Publishing. ISBN 978-3-030-10997-4.

Dimitris Ballas, Graham Clarke, Danny Dorling, Heather Eyre, Bethan Thomas, and David Rossiter. Simbritain: A spatial microsimulation approach to population dynamics. *Population, Space and Place*, 11:13 – 34, 01 2005. doi: 10.1002/psp.351.

Haris Ballis and Loukas Dimitriou. Revealing personal activities schedules from synthesizing multi-period origin-destination matrices. *Transportation Research Part B Methodological*, 139:224–258, 07 2020.

Maram Bani Younes and Azzedine Boukerche. Intelligent traffic light controlling algorithms using vehicular networks. *IEEE Transactions on Vehicular Technology*, 65(8):5887–5899, 2016. doi: 10.1109/TVT.2015.2472367.

Hillel Bar-Gera, Karthik C. Konduri, Bhargava Sana, Xin Ye, and Ram M. Pendyala. Estimating survey weights with multiple constraints using entropy optimization methods. In *88th Annual Meeting of the Transportation Research Board*, 2009.

Jaume Barceló and Lídia Montero. A robust framework for the estimation of dynamic od trip matrices for reliable traffic management. *Transportation Research Procedia*, 10:134–144, 2015. ISSN 2352-1465. doi: https://doi.org/10.1016/j.trpro.2015.09.063. URL https://www.sciencedirect.com/science/article/pii/S2352146515002501. 18th Euro Working Group on Transportation, EWGT 2015, 14-16 July 2015, Delft, The Netherlands.

Richard Beckman, Keith Baggerly, and Michael McKay. Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, 30:415–429, 11 1996.

Martin Beckmann, Charles B McGuire, and Christopher B Winsten. *Studies in the Economics of Transportation*. Research memorandum. Rand Corporation, 1955. URL https://books.google.fr/books?id=gbUeAQAAMAAJ.

Robert B. Bendel, Sahran S. Higgins, J. E. Teberg, and David A. Pyke. Comparison of skewness coefficient, coefficient of variation, and gini coefficient as inequality measures within populations. *Oecologia*, 78(3): 394–400, 1989.

Mounir Bensalah, Abdelmajid Elouadi, and Hassan Mharzi. OPTIMIZATION OF COST OF A TRAM THROUGH THE INTEGRATION OF BIM: A THEORETICAL ANALYSIS. *International Journal of Mechanical And Production Engineering*, 5(11):138 – 142, November 2017. URL https://hal.science/hal-02146976.

Claudio Bettini, X. Sean Wang, and Sushil Jajodia. Protecting privacy against location-based personal identification. In Willem Jonker and Milan Petković, editors, *Secure Data Management*, pages 185–199, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31974-0.

Tomas Beuzen, Lucy Marshall, and Kristen D Splinter. A comparison of methods for discretizing continuous variables in bayesian networks. *Environmental modelling & software*, 108:61–66, 2018.

Chandra R. Bhat, Jay P. Carini, and Rajul Misra. Modeling the generation and organization of household activity stops. *Transportation Research Record*, 1676(1):153–161, 1999. doi: 10.3141/1676-19. URL https://doi.org/10.3141/1676-19.

Michel Bierlaire and Frank Crittin. An efficient algorithm for real-time estimation and prediction of dynamic od tables. *Operations Research*, 52(1):116–127, 2004. ISSN 0030364X, 15265463. URL http://www.jstor.org/stable/30036564.

Federico Bigi, Nicola Schwemmle, and Francesco Viti. Evaluating the impact of free public transport using agent-based modeling: the case-study of luxembourg. In *11th Symposium of the European Association for Research in Transportation (hEART2023)*, 09 2023.

Vincent Bindschaedler and Reza Shokri. Synthesizing plausible privacy-preserving location traces. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 546–563, May 2016. doi: 10.1109/SP.2016.39.

Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Yvonne Millicent Mahala Bishop. *Multidimensional Contingency Tables: Cell Estimates*. ProQuest LLC, Ann Arbor, MI, 1967. URL http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&res_dat=xri:pqdiss&rft_dat=xri:pqdiss:7409260. Thesis (Ph.D.)–Harvard University.

Vincent D Blondel, Markus Esch, Connie Chan, Fabrice Clérot, Pierre Deville, Etienne Huens, Frédéric Morlot, Zbigniew Smoreda, and Cezary Ziemlicki. Data for development: the D4D challenge on mobile phone data. *ArXiv*, abs/1210.0137v2, 2013.

Otto Anker Nielsen Bo Friis Nielsen, Laura Frølich and Dorte Filges. Estimating passenger numbers in trains using existing weighing capabilities. *Transportmetrica A: Transport Science*, 10(6):502–517, 2014. doi: 10.1080/23249935.2013.795199. URL https://doi.org/10.1080/23249935.2013.795199.

Loic Bonnetain. *Unlocking the potential of mobile phone data for large scale urban mobility estimation*. Theses, Université de Lyon, February 2022. URL https://theses.hal.science/tel-03920673.

Loïc Bonnetain, Angelo Furno, Jean Krug, and Nour-Eddin El Faouzi. Can we map-match individual cellular network signaling trajectories in urban environments? data-driven study. *Transportation Research Record*, 2673(7):74–88, 2019. doi: 10.1177/0361198119847472. URL https://doi.org/10.1177/0361198119847472.

Loïc Bonnetain, Angelo Furno, Nour-Eddin El Faouzi, Marco Fiore, Razvan Stanica, Zbigniew Smoreda, and Cezary Ziemlicki. Transit: Fine-grained human mobility trajectory inference at scale with mobile network signaling data. *Transportation Research Part C: Emerging Technologies*, 130:103257, 2021. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2021.103257. URL https://www.sciencedirect.com/science/article/pii/S0968090X21002692.

Peter W. Bonsall. Chapter 5 - principles of transport analysis and forecasting. In CA O'Flaherty, MGH Bell, PW Bonsall, GR Leake, AD May, CA Nash, and CA O'Flaherty, editors, *Transport Planning and Traffic Engineering*, pages 103–131. Butterworth-Heinemann, Oxford, 1997. ISBN 978-0-340-66279-3. doi: https://doi.org/10.1016/B978-034066279-3/50007-4. URL https://www.sciencedirect.com/science/article/pii/B9780340662793500074.

Stanislav S. Borysov, Jeppe Rich, and Francisco C. Pereira. How to generate micro-agents? a deep generative modeling approach to population synthesis. *Transportation Research Part C: Emerging Technologies*, 106:73–97, 2019. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2019.07.006. URL https://www.sciencedirect.com/science/article/pii/S0968090X1831180X.

J.L Bowman and M.E Ben-Akiva. Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A: Policy and Practice*, 35(1):1–28, 2001. ISSN 0965-8564. doi: https://doi.org/10.1016/S0965-8564(99)00043-9. URL https://www.sciencedirect.com/science/article/pii/S0965856499000439.

Hamparsum Bozdogan. Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, 52(3):345–370, 1987.

Charlotte Brannigan, Marius Biedka, Guy Hitchcock, and Davide Fiorello TRT. Study on urban mobility—assessing and improving the accessibility of urban areas final report and policy proposals. https://docs.confebus.org/CE_MovilidadUrbana_MI-04-16-271-EN-N.pdf, 2017. [Online; accessed 24-August-2023].

Elmar Brockfeld, Peter Wagner, Stefan Lorkowski, and Peter Mieth. Benefits and limits of recent floating car data technology – an evaluation study. In *CDROM-WCTR2007*, pages C2–830, June 2007. URL https://elib.dlr.de/49618/.

Martin Brosnan, Michael Petesch, Jason Pieper, Scott Schumacher, and Greg Lindsey. Validation of bicycle counts from pneumatic tube counters in mixed traffic flows. *Transportation Research Record*, 2527(1):80–89, 2015. doi: 10.3141/2527-09. URL https://doi.org/10.3141/2527-09.

Andrew Bwambale, Charisma F. Choudhury, and Stephane Hess. Modelling trip generation using mobile phone data: A latent demographics approach. *Journal of Transport Geography*, 76:276–286, 2019. ISSN 0966-6923. doi: https://doi.org/10.1016/j.jtrangeo.2017.08.020. URL https://www.sciencedirect.com/science/article/pii/S0966692317301382.

Andrew Bwambale, Charisma F. Choudhury, Stephane Hess, and Md. Shahadat Iqbal. Getting the best of both worlds: a framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling. *Transportation*, 48(5):2287–2314, 2021. doi: 10.1007/s11116-020-10129-5. URL https://doi.org/10.1007/s11116-020-10129-5.

Jongik Byun and Roussos Dimitrakopoulos. An efficient algorithm for the lp relaxation of the maximal closure problem with a capacity constraint. Technical Report G–2013–60, Groupe d'études et de recherche en analyse des décisions, GERAD, Montréal QC H3T 2A7, Canada, 2013. URL https://www.gerad.ca/en/papers/G-2013-60.

Sally Cairns, Stephen Atkins, and Phil Goodwin. Disappearing traffic? the story so far. *Proceedings of the Institution of Civil Engineers - Municipal Engineer*, 151(1):13–22, 2002. doi: 10.1680/muen.2002.151.1.13. URL https://doi.org/10.1680/muen.2002.151.1.13.

Paul H. Calamai and Jorge J. Moré. Projected gradient methods for linearly constrained problems. *Mathematical Programming*, 39(1):93–116, 1987.

Guido Cantelmo, Francesco Viti, Ernesto Cipriani, and Nigro Marialisa. A two-steps dynamic demand estimation approach sequentially adjusting generations and distributions. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 1477–1482, 2015. doi: 10.1109/ITSC.2015.241.

Chris K Carter and Robert Kohn. On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553, 09 1994. ISSN 0006-3444. doi: 10.1093/biomet/81.3.541. URL https://doi.org/10.1093/biomet/81.3.541.

Ennio Cascetta and Sang Nguyen. A unified framework for estimating or updating origin/destination matrices from traffic counts. *Transportation Research Part B: Methodological*, 22(6):437–455, 1988. ISSN 0191-2615. doi: https://doi.org/10.1016/0191-2615(88)90024-0. URL https://www.sciencedirect.com/science/article/pii/0191261588900240.

Ennio Cascetta, Andrea Papola, Vittorio Marzano, Fulvio Simonelli, and Iolanda Vitiello. Quasi-dynamic estimation of o–d flows from traffic counts: Formulation, statistical validation and performance analysis on real data. *Transportation Research Part B: Methodological*, 55:171–187, 2013. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2013.06.007. URL https://www.sciencedirect.com/science/article/pii/S0191261513001069.

Cerema. lil-1023: Enquête ménage déplacement, lyon / aire métropolitaine lyonnaise. https://data.progedo.fr/studies/doi/10.13144/lil-1023, 2015.

CEREMA. Enquête ménages déplacements, lyon / aire métropolitaine lyonnaise (emd, lyon / aire métropolitaine lyonnaise). https://data.progedo.fr/studies/doi/10.13144/lil-1023, 2015. Syndicat mixte des transports pour le Rhône et l'agglomération lyonnaise (producer), ADISP (distributor).

Vitor Cerqueira, Luis Moreira-Matias, Jihed Khiari, and Hans van Lint. On evaluating floating car data quality for knowledge discovery. *IEEE Transactions on Intelligent Transportation Systems*, 19(11):3749–3760, 2018. doi: 10.1109/TITS.2018.2867834.

Kevin Chapuis, Patrick Taillandier, Renaud Misslin, and Alexis Drogoul. Gen*: a generic toolkit to generate spatially explicit synthetic populations. *International Journal of Geographical Information Science*, 32 (6):1194–1210, 2018. doi: 10.1080/13658816.2018.1440563. URL https://hal.sorbonne-universite.fr/hal-01821597.

Bi Yu Chen, Yafei Wang, Donggen Wang, Qingquan Li, William Lam, and Shih-Lung Shaw. Understanding the impacts of human mobility on accessibility using massive mobile phone tracking data. *Annals of the Association of American Geographers*, 108:1115–1133, 01 2018. doi: 10.1080/24694452.2017.1411244.

Lei Chen, M. Tamer Özsu, and Vincent Oria. Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, page 491–502, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595930604. doi: 10.1145/1066157.1066213. URL https://doi.org/10.1145/1066157.1066213.

Xumei Chen, Shuxia Guo, Lei Yu, and Bruce Hellinga. Short-term forecasting of transit route od matrix with smart card data. In *Conference Record - IEEE Conference on Intelligent Transportation Systems*, pages 1513–1518, 10 2011. ISBN 978-1-4577-2198-4. doi: 10.1109/ITSC.2011.6082929.

Long Cheng, Xuewu Chen, and Shuo Yang. An exploration of the relationships between socioeconomics, land use and daily trip chain pattern among low-income residents. *Transportation Planning and Technology*, 39 (4):358–369, 2016.

Stefano Chiesa and Sergio Taraglio. Traffic request generation through a variational auto encoder approach. *Computers*, 11(5):71, 2022.

Julie Chrétien, Florent Le Néchet, Fabien Leurent, and Biao Yin. Using mobile phone data to observe and understand mobility behavior, territories, and transport usage. *Urban Mobility and the Smartphone: Transportation, Travel Behavior and Public Policy*, 79, 2018.

CNIS. https://www.cnis.fr/wp-content/uploads/2018/01/AE_2017_IDF_MobilitÃ¯s_EGT.pdf, 2018. [Online; accessed 9-October-2023].

Barry R. Cobb, Rafael Rumí, and Antonio Salmerón. *Bayesian Network Models with Discrete and Continuous Variables*, pages 81–102. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007. ISBN 978-3-540-68996-6. doi: 10.1007/978-3-540-68996-6_4. URL https://doi.org/10.1007/978-3-540-68996-6_4.

Commission des finances. Grand paris express: des coûts à maîtriser, un financement à consolider, 2020. URL https://www.senat.fr/rap/r20-044/r20-04413.html.

Graham Cormode, Magda Procopiuc, Divesh Srivastava, and Thanh Tran. Differentially private publication of sparse data. In *ICDT '12: Proceedings of the 15th International Conference on Database Theory*, pages 299–311, 03 2011. doi: 10.1145/2274576.2274608.

Antonello Ignazio Croce, Giuseppe Musolino, Corrado Rindone, and Antonino Vitetta. Estimation of travel demand models with limited information: Floating car data for parameters' calibration. *Sustainability*, 13 (16), 2021. ISSN 2071-1050. doi: 10.3390/su13168838. URL https://www.mdpi.com/2071-1050/13/16/8838.

George B. Dantzig. *Origins of the Simplex Method*, page 141–151. Association for Computing Machinery, New York, NY, USA, 1990. ISBN 0201508141. URL https://doi.org/10.1145/87252.88081.

Gareth Davies. *Examination of approaches to calibration in survey sampling*. PhD thesis, Cardiff University, 2018.

Yves-Alexandre de Montjoye, Zbigniew Smoreda, Romain Trinquart, Cezary Ziemlicki, and Vincent D. Blondel. D4d-senegal: The second mobile phone data for development challenge. *ArXiv*, abs/arXiv:1407.4885, 2014.

Jean-Claude Deville, Carl-Erik Sarndal, and Olivier Sautory. Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88(423):1013–1020, 1993. ISSN 01621459. URL http://www.jstor.org/stable/2290793.

Felipe F. Dias, Gopindra S. Nair, Natalia Ruíz-Juri, Chandra R. Bhat, and Arash Mirzaei. Incorporating autonomous vehicles in the traditional four-step model. *Transportation Research Record*, 2674(7):348–360, 2020. doi: 10.1177/0361198120922544. URL https://doi.org/10.1177/0361198120922544.

Katerina Doka, Mingqiang Xue, Dimitrios Tsoumakos, and Panagiotis Karras. K-anonymization by freeform generalization. In *Proceedings of the 10th ACM Symposium on Information, Computer and Communications Security*, ASIA CCS '15, page 519–530, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450332453. doi: 10.1145/2714576.2714590. URL https://doi.org/10.1145/2714576.2714590.

Marcello D'Orazio, Marco Zio, and Mauro Scanu. *Statistical Matching: Theory and Practice*. John Wiley & Sons, 04 2006. ISBN 0-470-02353-8. doi: 10.1002/0470023554.

Cynthia Dwork. A firm foundation for private data analysis. *Commun. ACM*, 54(1):86–95, January 2011. ISSN 0001-0782. doi: 10.1145/1866739.1866758. URL https://doi.org/10.1145/1866739.1866758.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, August 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-32732-5.

Oscar Egu and Patrick Bonnel. How comparable are origin-destination matrices estimated from automatic fare collection, origin-destination surveys and household travel survey? an empirical investigation in lyon. *Transportation Research Part A: Policy and Practice*, 138:267–282, 2020. ISSN 0965-8564. doi: https://doi.org/10.1016/j.tra.2020.05.021. URL https://www.sciencedirect.com/science/article/pii/S0965856420306030.

European Commission. 2018 reform of EU data protection rules. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32016R0679, 2018.

Digital cellular telecommunications system (Phase 2+); Handover procedures. European Telecommunications Standards Institute, 8 1997.

Universal Mobile Telecommunications System (UMTS); Handover procedures. European Telecommunications Standards Institute, 3 2001.

Bilal Farooq, Michel Bierlaire, Ricardo Hurtubia, and Gunnar Flötteröd. Simulation based population synthesis. *Transportation Research Part B: Methodological*, 58, 12 2013.

Antonio Fernández, Inmaculada Pérez-Bernabé, and Antonio Salmerón. On using the pc algorithm for learning continuous bayesian networks: An experimental analysis. In Concha Bielza, Antonio Salmerón, Amparo Alonso-Betanzos, J. Ignacio Hidalgo, Luis Martínez, Alicia Troncoso, Emilio Corchado, and Juan M. Corchado, editors, *Advances in Artificial Intelligence*, pages 342–351, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40643-0.

Winnie Daamen Florian Schneider, Anders Fjendbo Jensen and Serge Hoogendoorn. Empirical analysis of cycling distances in three of europe's most bicycle-friendly regions within an accessibility framework. *International Journal of Sustainable Transportation*, 17(7):775–789, 2023. doi: 10.1080/15568318.2022.2095945. URL https://doi.org/10.1080/15568318.2022.2095945.

Mohammad Forghani, Farid Karimipour, and Christophe Claramunt. From cellular positioning data to trajectories: Steps towards a more accurate mobility exploration. *Transportation Research Part C: Emerging Technologies*, 117:102666, 2020. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2020.102666. URL https://www.sciencedirect.com/science/article/pii/S0968090X20305817.

Nicholas Fournier, Eleni Christofa, Arun Akkinepally, and Carlos Lima Azevedo. Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method. *Transportation*, 48, 04 2021.

Rodric Frederix, Francesco Viti, Willem WE Himpe, and Chris MJ Tampère. Dynamic origin–destination matrix estimation on large-scale congested networks using a hierarchical decomposition scheme. *Journal of Intelligent Transportation Systems*, 18(1):51–66, 2014.

Jerome Friedman, Jon Bentley, and Raphael Finkel. An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Softw.*, 3:209–226, 09 1977. doi: 10.1145/355744.355745.

Marcia R Friesen and Robert D McLeod. Bluetooth in intelligent transportation systems: a survey. *International Journal of Intelligent Transportation Systems Research*, 13:143–153, 2015.

Benjamin C. M. Fung, Ke Wang, Rui Chen, and Philip S. Yu. Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, 42(4), jun 2010. ISSN 0360-0300. doi: 10.1145/1749603.1749605. URL https://doi.org/10.1145/1749603.1749605.

Emanuele Galli, Leticia Cuellar, Stephan Eidenbenz, Mary Ewers, Susan Mniszewski, and Christof Teuscher. Activitysim: large-scale agent-based activity generation for infrastructure simulation. In *Proceedings of the 2009 Spring Simulation Multiconference, SpringSim 2009*, 01 2009. doi: 10.1145/1639809.1639826.

Sergio Garrido, Stanislav Borysov, Francisco Pereira, and J. Rich. Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C: Emerging Technologies*, 120:102787, 11 2020.

Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *6th proceedings of the IEEE Transactions on pattern analysis and machine intelligence*, pages 721–741, 1984.

Arthur M Geoffrion. Lagrangean relaxation for integer programming. In *Approaches to integer programming*, pages 82–114. Springer, 2009.

Suxia Gong, Ismaïl Saadi, Jacques Teller, and Mario Cools. How comparable are mobile phone data and travel survey data? In *BIVEC-GIBET Transport Research Days 2023*, 2023.

Andrea Gorrini, Luca Crociani, Giuseppe Vizzari, and Stefania Bandini. Observation results on pedestrian-vehicle interactions at non-signalized intersections towards simulation. *Transportation Research Part F: Traffic Psychology and Behaviour*, 59:269–285, 2018. ISSN 1369-8478. doi: https://doi.org/10.1016/j.trf.2018.09.016. URL https://www.sciencedirect.com/science/article/pii/S1369847817302000.

Eduardo Graells-Garrido, Diego Caro, and Denis Parra. Inferring modes of transportation using mobile phone data. *EPJ Data Science*, 7(1):49, 2018. doi: 10.1140/epjds/s13688-018-0177-1. URL https://doi.org/10.1140/epjds/s13688-018-0177-1.

Marco Gramaglia and Marco Fiore. Hiding mobile traffic fingerprints with glove. In *Proceedings of the 11th ACM Conference on Emerging Networking Experiments and Technologies*, CoNEXT '15, pages 1–13, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334129. doi: 10.1145/2716281.2836111. URL https://doi.org/10.1145/2716281.2836111.

Marco Gramaglia, Marco Fiore, Alberto Tarable, and Albert Banchs. Preserving mobile subscriber privacy in open datasets of spatiotemporal trajectories. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pages 1–9, 05 2017. doi: 10.1109/INFOCOM.2017.8056979.

Fangce Guo. *Short-term traffic prediction under normal and abnormal conditions*. PhD thesis, Imperial College London, 2013.

Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 1737–1746. JMLR.org, 2015.

Alex Hagen-Zanker and Ying Jin. Adaptive zoning for efficient transport modelling in urban models. In Osvaldo Gervasi, Beniamino Murgante, Sanjay Misra, Marina L. Gavrilova, Ana Maria Alves Coutinho Rocha, Carmelo Torre, David Taniar, and Bernady O. Apduhan, editors, *Computational Science and Its Applications – ICCSA 2015*, pages 673–687, Cham, 2015. Springer International Publishing. ISBN 978-3-319-21470-2.

Hayati Hasibuan and Mari Mulyani. Transit-oriented development: Towards achieving sustainable transport and urban development in jakarta metropolitan, indonesia. *Sustainability*, 14:5244, 04 2022. doi: 10.3390/su14095244.

Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014. doi: 10.1080/15230406.2014.890072. URL https://doi.org/10.1080/15230406.2014.890072.

Brian He, Jinkai Zhou, Ziyi Ma, Joseph Chow, and Kaan Ozbay. Evaluation of city-scale built environment policies in new york city with an emerging-mobility-accessible synthetic population. *Transportation Research Part A Policy and Practice*, 141:444–467, 10 2020.

Matthias Hess. Capital construction costs in urban subway systems: A comparison of two projects in london and new york, 2016.

Tim Hilgert, Michael Heilig, Martin Kagerbauer, and Peter Vortisch. Modeling week activity schedules for travel demand models. *Transportation Research Record*, 2666(1):69–77, 2017. doi: 10.3141/2666-08. URL https://doi.org/10.3141/2666-08.

Julian Holzel. Differential privacy and the gdpr. *European Data Protection Law Review*, 5:184–196, 01 2019. doi: 10.21552/edpl/2019/2/8.

Sebastian Hörl and Kay Axhausen. Relaxation–discretization algorithm for spatially constrained secondary location assignment. *Transportmetrica A: Transport Science*, 10 2021. doi: 10.1080/23249935.2021.1982068.

Sebastian Hörl and Milos Balac. Synthetic population and travel demand for paris and île-de-france based on open and publicly available data. *Transportation Research Part C: Emerging Technologies*, 130:103291, 2021. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2021.103291. URL https://www.sciencedirect.com/science/article/pii/S0968090X21003016.

Andreas Horni, Kai Nagel, and Kay Axhausen. *The Multi-Agent Transport Simulation MATSim*. Ubiquity Press, 04 2016.

Arthur Huang and David Levinson. Axis of travel: Modeling non-work destination choice with gps data. *Transportation Research Part C: Emerging Technologies*, 58, 04 2015. doi: 10.1016/j.trc.2015.03.022.

Zhiren Huang, Ximan Ling, Pu Wang, Fan Zhang, Yingping Mao, and Tao Lin. Modeling real-time human mobility based on mobile phone and transportation data fusion. *Transportation Research Part C Emerging Technologies*, 96:251–269, 10 2018.

Perry Hystad, Ofer Amram, Funso Oje, Andrew Larkin, Kwadwo Boakye, Ally Avery, Assefaw Gebremedhin, and Glen Duncan. Bring your own location data: Use of google smartphone location history data for environmental health research. *Environmental Health Perspectives*, 130(11):117005, 2022. doi: 10.1289/EHP10829. URL https://ehp.niehs.nih.gov/doi/abs/10.1289/EHP10829.

Anugrah Ilahi and Kay Axhausen. Integrating bayesian network and generalized raking for population synthesis in greater jakarta. *Regional Studies, Regional Science*, 6, 11 2019.

INSEE. Professions et catégories socioprofessionnelles. https://www.insee.fr/fr/metadonnees/pcs2003/categorieSocioprofessionnelleAgregee/1, 2003. Accessed: 2022-10-28.

INSEE. https://www.insee.fr/fr/statistiques/3625223, 2018. [Online; accessed 30-November-2023].

INSEE. Mobilités professionnelles en 2007 : déplacements domicile - lieu de travail. https://www.insee.fr/fr/statistiques/2022121, 2020.

Md Shahadat Iqbal, Charisma Choudhury, Pu Wang, and Marta C. Gonzalez. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63–74, 03 2014.

Walter Isard. Location Theory and Trade Theory: Short-Run Analysis. *The Quarterly Journal of Economics*, 68(2):305–320, 05 1954. ISSN 0033-5533. doi: 10.2307/1884452. URL https://doi.org/10.2307/1884452.

Frank Jensen, Finn V Jensen, and Søren L Dittmer. From influence diagrams to junction trees. In *Uncertainty Proceedings 1994*, pages 367–373. Elsevier, 1994.

Kaifeng Jiang, Dongxu Shao, Stéphane Bressan, Thomas Kister, and Kian-Lee Tan. Publishing trajectories with differential privacy guarantees. In *Proceedings of the 25th International Conference on Scientific and Statistical Database Management*, SSDBM, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450319218. doi: 10.1145/2484838.2484846. URL https://doi.org/10.1145/2484838.2484846.

Shan Jiang, Yingxiang Yang, Siddharth Gupta, Daniele Veneziano, Shounak Athavale, and Marta C. Gonzalez. The timegeo modeling framework for urban mobility without travel surveys. *Proceedings of the National Academy of Sciences*, 113:201524261, 08 2016. doi: 10.1073/pnas.1524261113.

Shan Jiang, Joseph Ferreira, and Marta C. Gonzalez. Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. *IEEE Transactions on Big Data*, 3(2):208–219, 2017. doi: 10.1109/TBDATA.2016.2631141.

Fengmei Jin, Wen Hua, Boyu Ruan, and Xiaofang Zhou. Frequency-based randomization for guaranteeing differential privacy in spatial trajectories, 2022. URL https://arxiv.org/abs/2207.03722.

Martin Johnsen, Oliver Brandt, Sergio Garrido, and Francisco Pereira. Population synthesis for urban resident modeling using deep generative models. *Neural Computing and Applications*, 34, 03 2022. doi: 10.1007/s00521-021-06622-2.

David S. Johnson and K.A. Niemi. On knapsacks, partitions, and a new dynamic programming technique for trees. *Math. Oper. Res.*, 8(1):1–14, February 1983. ISSN 0364-765X. doi: 10.1287/moor.8.1.1. URL https://doi.org/10.1287/moor.8.1.1.

Michael Jordan. An introduction to graphical models. unpublished, 2015. URL http://people.eecs.berkeley.edu/~jordan/prelims/.

Johan W Joubert and Alta De Waal. Activity-based travel demand generation using bayesian networks. *Transportation Research Part C Emerging Technologies*, 120, 09 2020.

Andreas Justen, Francisco J. Martínez, and Cristián E. Cortés. The use of space–time constraints for the selection of discretionary activity locations. *Journal of Transport Geography*, 33:146–152, 2013. ISSN 0966-6923. doi: https://doi.org/10.1016/j.jtrangeo.2013.10.009. URL https://www.sciencedirect.com/science/article/pii/S0966692313002068.

Mahfuz Kabir, Ruhul Salim, and Nasser Al-Mawali. The gravity model and trade flows: Recent developments in econometric modeling and empirical evidence. *Economic Analysis and Policy*, 56:60–71, 2017. ISSN 0313-5926. doi: https://doi.org/10.1016/j.eap.2017.08.005. URL https://www.sciencedirect.com/science/article/pii/S0313592617300899.

Alexandra Kapp, Julia Hansmeyer, and Helena Mihaljević. Generative models for synthetic urban mobility data: A systematic literature review. *ACM Comput. Surv.*, 56(4), nov 2023. ISSN 0360-0300. doi: 10.1145/3610224. URL https://doi.org/10.1145/3610224.

Narendra Karmarkar. An interior-point approach to NP-complete problems. I. In *Mathematical developments arising from linear programming (Brunswick, ME, 1988)*, volume 114 of *Contemp. Math.*, pages 297–308. Amer. Math. Soc., Providence, RI, 1990. ISBN 0-8218-5121-7. doi: 10.1090/conm/114/1097880. URL https://doi.org/10.1090/conm/114/1097880.

Ankit Kathuria and Perumal Vedagiri. Evaluating pedestrian vehicle interaction dynamics at un-signalized intersections: A proactive approach for safety analysis. *Accident Analysis & Prevention*, 134:105316, 2020. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2019.105316. URL https://www.sciencedirect.com/science/article/pii/S0001457519303847.

Hans Kellerer, Ulrich Pferschy, and David Pisinger. *Knapsack Problems*. Springer, first edition, 01 2004. ISBN 978-3-540-40286-2. doi: 10.1007/978-3-540-24777-7.

Boris S. Kerner, C. Demir, R.G. Herrtwich, S.L. Klenov, H. Rehborn, M. Aleksic, and A. Haug. Traffic state detection with floating car data in road networks. In *Proceedings. 2005 IEEE Intelligent Transportation Systems, 2005.*, pages 44–49, 2005. doi: 10.1109/ITSC.2005.1520133.

Leonid G Khachiyan. Polynomial algorithms in linear programming. *USSR Computational Mathematics and Mathematical Physics*, 20(1):53–72, 1980. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(80)90061-0. URL https://www.sciencedirect.com/science/article/pii/0041555380900610.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2014.

Karthik Konduri, Daehyun You, Venu Garikapati, and Ram Pendyala. Enhanced synthetic population generator that accommodates control variables at multiple geographic resolutions. *Transportation Research Record: Journal of the Transportation Research Board*, 2563:40–50, 01 2016. doi: 10.3141/2563-08.

Victor Kyriacou, Yiolanda Englezou, Christos G. Panayiotou, and Stelios Timotheou. Bayesian traffic state estimation using extended floating car data. *IEEE Transactions on Intelligent Transportation Systems*, 24 (2):1518–1532, 2023. doi: 10.1109/TITS.2022.3225057.

Pierre-Antoine Laharotte, Romain Billot, Etienne Come, Latifa Oukhellou, Alfredo Nantes, and Nour-Eddin El Faouzi. Spatiotemporal analysis of bluetooth data: Application to a large urban network. *IEEE Transactions on Intelligent Transportation Systems*, 16(3):1439–1448, 2015. doi: 10.1109/TITS.2014.2367165.

Jeffrey Larson, Matt Menickelly, and Stefan M. Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019. doi: 10.1017/S0962492919000060.

Steffen Lauritzen and David Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *The Journal of the Royal Statistical Society. Series B (Methodological)*, 50:157–224, 01 1988.

Jean-Patrick Lebacque. Intersection modeling, application to macroscopic network traffic flow models and traffic management. In Serge P. Hoogendoorn, Stefan Luding, Piet H. L. Bovy, Michael Schreckenberg, and Dietrich E. Wolf, editors, *Traffic and Granular Flow '03*, pages 261–278, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-28091-0.

Kristen LeFevre, David J DeWitt, and Raghu Ramakrishnan. Mondrian multidimensional k-anonymity. In *Proceedings of the 22nd International Conference on Data Engineering*, volume 2006, pages 25 – 25, 05 2006. ISBN 0-7695-2570-9. doi: 10.1109/ICDE.2006.101.

Claude Lemaréchal. Lagrangian relaxation. *Computational combinatorial optimization: optimal or provably near-optimal solutions*, pages 112–156, 2001.

Bo Lenntorp. *Paths in Space-time Environments: A Time-geographic Study of Movement Possibilities of Individuals*. Geografiska Institution Lund: Meddelanden. Royal University of Lund, Department of Geography, 1976. ISBN 9789140043764. URL https://books.google.fr/books?id=haseAQAAMAAJ.

Maxime Lenormand, Miguel Picornell, Oliva G. Cantú-Ros, Antònia Tugores, Thomas Louail, Ricardo Herranz, Marc Barthelemy, Enrique Frías-Martínez, and José J. Ramasco. Cross-checking different sources of mobility information. *PLOS ONE*, 9(8):1–10, 08 2014. doi: 10.1371/journal.pone.0105184. URL https://doi.org/10.1371/journal.pone.0105184.

Asad Lesani and Luis Miranda-Moreno. Development and testing of a real-time wifi-bluetooth system for pedestrian network monitoring, classification, and data extrapolation. *IEEE Transactions on Intelligent Transportation Systems*, 20(4):1484–1496, 2019. doi: 10.1109/TITS.2018.2854895.

Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd International Conference on Data Engineering*, pages 106–115, 2007. doi: 10.1109/ICDE.2007.367856.

Stan Z. Li. *MRF Parameter Estimation*, pages 1–32. Springer London, London, 2009. ISBN 978-1-84800-279-1. doi: 10.1007/978-1-84800-279-1_7. URL https://doi.org/10.1007/978-1-84800-279-1_7.

Zhong-Xian Li, Xiao-Ming Yang, and Zongjin Li. Application of cement-based piezoelectric sensors for monitoring traffic flows. *Journal of transportation engineering*, 132(7):565–573, 2006.

Yuting Liang and Reza Samavi. Optimization-based k-anonymity algorithms. *Computers & Security*, 93 (101753):1–17, 2020. ISSN 0167-4048. doi: https://doi.org/10.1016/j.cose.2020.101753. URL https://www.sciencedirect.com/science/article/pii/S0167404820300377.

Yifan Liao. Hot spot analysis of tourist attractions based on stay point spatial clustering. *Journal of Information Processing Systems*, 16(4):750–759, 2020.

Ziheng Lin, Mogeng Yin, Sidney A. Feygin, Madeleine Sheehan, Jean-Franǫis Paiement, and Alexei Pozdnoukhov Cee. Deep generative models of urban mobility. In *IEEE Transactions on Intelligent Transportation Systems*, 2017.

Dennis Lindley. Kendall's Advanced Theory of Statistics, volume 2B, Bayesian Inference, 2nd edn. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 168(1):259–260, 12 2004. ISSN 0964-1998. doi: 10.1111/j.1467-985X.2004.00347_15.x. URL https://doi.org/10.1111/j.1467-985X.2004.00347_15.x.

Er-Jian Liu and Xiao-Yong Yan. A universal opportunity model for human mobility. *Scientific Reports*, 10: 4657, 03 2020.

Feng Liu, Davy Janssens, Jianxun Cui, Geert Wets, and Mario Cools. Characterizing activity sequences using profile hidden markov models. *Expert Systems with Applications*, 42, 03 2015.

Xingjian Liu and Ying Long. Automated identification and characterization of parcels with openstreetmap and points of interest. *Environment and Planning B: Planning and Design*, 43(2):341–360, 2016. doi: 10.1177/0265813515604767. URL https://doi.org/10.1177/0265813515604767.

Max O Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905. ISSN 15225437. URL http://www.jstor.org/stable/2276207.

Robin Lovelace, Mark Birkin, Dimitris Ballas, and Eveline S. Leeuwen. Evaluating the performance of iterative proportional fitting for spatial microsimulation: New tests for an established technique. *Journal of Artificial Societies and Social Simulation, The*, 18:21, 03 2015.

Tai-yu Ma and Sylvain Klein. Bayesian networks for constrained location choice modeling using structural restrictions and model averaging. *European Journal of Transport and Infrastructure Research*, 18, 09 2017.

Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. L-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data*, 1(1):3–15, March 2007. ISSN 1556-4681. doi: 10.1145/1217299.1217302. URL https://doi.org/10.1145/1217299.1217302.

Waranya Mahanan, W. Chaovalitwongse, and Juggapong Natwichai. Data privacy preservation algorithm with k-anonymity. *World Wide Web*, 24:1551–1561, 09 2021. doi: 10.1007/s11280-021-00922-2.

Michael J Maher. Inferences on trip matrices from observations on link volumes: A bayesian statistical approach. *Transportation Research Part B: Methodological*, 17(6):435–447, 1983. ISSN 0191-2615. doi: https://doi.org/10.1016/0191-2615(83)90030-9. URL https://www.sciencedirect.com/science/article/pii/0191261583900309.

Nayan Maiti, Pranjal Pathak, and Biswajit Samanta. An efficient algorithm for the precedence constraint knapsack problem with reference to large-scale open-pit mining pushback design. *Mining Technology*, 130: 1–14, 01 2021. doi: 10.1080/25726668.2020.1866369.

Marco Mamei, Nicola Bicocchi, Marco Lippi, Stefano Mariani, and Franco Zambonelli. Evaluating origin–destination matrices obtained from cdr data. *Sensors*, 19(20), 2019. ISSN 1424-8220. doi: 10.3390/s19204470. URL https://www.mdpi.com/1424-8220/19/20/4470.

Orlando Martínez-Durive, Theo Couturieux, Cezary Ziemlicki, and Marco Fiore. Voronoiboost: Data-driven probabilistic spatial mapping of mobile network metadata. In *2022 19th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 100–108, 2022. doi: 10.1109/SECON55815.2022.9918610.

Benoit Matet, Angelo Furno, Marco Fiore, Etienne Côme, and Latifa Oukhellou. Adaptative generalisation over a value hierarchy for the k-anonymisation of origin–destination matrices. *Transportation Research Part C: Emerging Technologies*, 154:104236, 2023a. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2023.104236. URL https://www.sciencedirect.com/science/article/pii/S0968090X23002255.

Benoît Matet, Etienne Côme, Angelo Furno, Loïc Bonnetain, Latifa Oukhellou, and Nour-Eddin El Faouzi. A lightweight approach for origin-destination matrix anonymization. In *29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 487–492, 01 2021a. doi: 10.14428/esann/2021.ES2021-56.

Benoît Matet, Etienne Côme, Angelo Furno, Latifa Oukhellou, and Nour-Eddin El Faouzi. Mobile phone origine-destination matrix anonymization and analysis. In *Proceedings of the Conference of Complex Systems*, page 130, 10 2021b. URL http://sci-web.net/CCS2021/CCS2021_BookOfAbstracts.pdf.

Benoît Matet, Etienne Côme, Angelo Furno, Sebastian Hörl, and Latifa Oukhellou. Use of origin-destination data in synthetic travel demand synthesis. In *10th International Symposium on Transportation Data & Modelling (ISTDM2023)*, pages 23–26, 06 2023b. doi: 10.2760/135735.

Benoît Matet, Etienne Côme, Angelo Furno, Sebastian Hörl, and Latifa Oukhellou. Use of origin-destination data for calibration and spatialization of synthetic travel demand. In *11th symposium of the European Association for Research in Transportation*, 09 2023c. URL https://transp-or.epfl.ch/heart/2023/abstracts/hEART_2023_paper_4604.pdf.

Benoît Matet, Etienne Côme, Angelo Furno, Sebastian Hörl, Latifa Oukhellou, and Nour-Eddin El Faouzi. Improving travel demand synthesis using origin-destination matrices from mobile phone data. under submission, 11 2024.

Nancy McGuckin and Elaine Murakami. Examining trip-chaining behavior: Comparison of travel by men and women. *Transportation Research Record*, 1693(1):79–85, 1999. doi: 10.3141/1693-12. URL https://doi.org/10.3141/1693-12.

Nicholas Monath, Kumar Avinava Dubey, Guru Guruganesh, Manzil Zaheer, Amr Ahmed, Andrew McCallum, Gokhan Mergen, Marc Najork, Mert Terzihan, Bryon Tjanaka, Yuan Wang, and Yuchen Wu. Scalable hierarchical agglomerative clustering. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1245–1255, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383325. doi: 10.1145/3447548.3467404. URL https://doi.org/10.1145/3447548.3467404.

Yves-Alexandre Montjoye, Cesar Hidalgo, Michel Verleysen, and Vincent D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3(1376):1–5, 03 2013. doi: 10.1038/srep01376.

Kirill Müller and Kay W. Axhausen. Population synthesis for microsimulation. state of the art. In *90th Annual Meeting of the Transportation Research Board*, volume 638, Zurich, 2010. IVT, ETH Zurich. doi: 10.3929/ethz-a-006127782.

D. Rahul Naik, Lyla B. Das, and T. S. Bindiya. Wireless sensor networks with zigbee and wifi for environment monitoring, traffic management and vehicle monitoring in smart cities. In *2018 IEEE 3rd International Conference on Computing, Communication and Security (ICCCS)*, pages 46–50, 2018. doi: 10.1109/CCCS.2018.8586819.

Bahar Namaki Araghi, Kristian Skoven Pedersen, Lars Tørholm Christensen, Rajesh Krishnan, and Harry Lahrmann. Accuracy of travel time estimation using bluetooth technology: Case study limfjord tunnel aalborg. *International Journal of Intelligent Transportation Systems Research*, 13(3):166–191, Sep 2015. ISSN 1868-8659. doi: 10.1007/s13177-014-0094-z. URL https://doi.org/10.1007/s13177-014-0094-z.

Mohammad-Reza Namazi-Rad, Robert Tanton, David Steel, Payam Mokhtarian, and Sumonkanti Das. An unconstrained statistical matching algorithm for combining individual and household level geo-specific census and survey data. *Computers, Environment and Urban Systems*, 63:3–14, 2017. ISSN 0198-9715. doi: https://doi.org/10.1016/j.compenvurbsys.2016.11.003. URL https://www.sciencedirect.com/science/article/pii/S0198971516303660. Spatial analysis with census data: emerging issues and innovative approaches.

Mohamad Belal Natafgi, Mohamad Osman, Asser Sleiman Haidar, and Lama Hamandi. Smart traffic light system using machine learning. In *2018 IEEE International Multidisciplinary Conference on Engineering Technology (IMCET)*, pages 1–6, 2018. doi: 10.1109/IMCET.2018.8603041.

Andrew A Neath and Joseph E Cavanaugh. The bayesian information criterion: background, derivation, and applications. *Wiley Interdisciplinary Reviews: Computational Statistics*, 4(2):199–203, 2012.

Samuel Nello-Deakin. Exploring traffic evaporation: Findings from tactical urbanism interventions in barcelona. *Case studies on transport policy*, 10(4):2430–2442, 2022.

James R Norris. *Markov chains*. Number 2 in Cambridge series on statistical and probabilistic mathematics. Cambridge university press, 1998.

Gregory Nuel. Tutorial on exact belief propagation in bayesian networks: from messages to algorithms, 2012.

Akande Noah Oluwatobi, Arulogun Oladiran Tayo, Aro Taye Oladele, and Ganiyu Rafiu Adesina. The design of a vehicle detector and counter system using inductive loop technology. *Procedia Computer Science*, 183:493–503, 2021. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2021.02.089. URL https://www.sciencedirect.com/science/article/pii/S1877050921005652. Proceedings of the 10th International Conference of Information and Communication Technology.

Kun Ouyang, Reza Shokri, David S. Rosenblum, and Wenzhuo Yang. A non-parametric generative model for human trajectories. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, page 3812–3817. AAAI Press, 2018. ISBN 9780999241127.

Teresa Pamuła and Renata Żochowska. Estimation and prediction of the od matrix in uncongested urban road network based on traffic flows using deep learning. *Engineering Applications of Artificial Intelligence*, 117:105550, 2023. ISSN 0952-1976. doi: https://doi.org/10.1016/j.engappai.2022.105550. URL https://www.sciencedirect.com/science/article/pii/S0952197622005401.

Luca Pappalardo and Filippo Simini. Data-driven generation of spatio-temporal routines in human mobility. *Data Mining and Knowledge Discovery*, 32, 05 2018. doi: 10.1007/s10618-017-0548-4.

Luca Pappalardo, Leo Ferres, Manuel Sacasa, Ciro Cattuto, and Loreto Bravo. Evaluation of home detection algorithms on mobile phone data using individual-level ground truth. *EPJ Data Science*, 10(1):29, 2021. doi: 10.1140/epjds/s13688-021-00284-9. URL https://doi.org/10.1140/epjds/s13688-021-00284-9.

Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 1558604790.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Jean-Claude Picard. Maximal closure of a graph and applications to combinatorial problems. *Management Science*, 22(11):1268–1272, 1976. ISSN 00251909, 15265501. URL http://www.jstor.org/stable/2630227.

Simon Porcher and Thomas Renault. Social distancing beliefs and human mobility: Evidence from twitter. *PLOS ONE*, 16(3):1–12, 03 2021. doi: 10.1371/journal.pone.0246949. URL https://doi.org/10.1371/journal.pone.0246949.

Janody Pougala, Tim Hillel, and Michel Bierlaire. Oasis: Optimisation-based activity scheduling with integrated simultaneous choice dimensions. *Transportation Research Part C: Emerging Technologies*, 155:104291, 2023. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2023.104291. URL https://www.sciencedirect.com/science/article/pii/S0968090X23002802.

Soora Rasouli and Harry Timmermans. Activity-based models of travel demand: Promises, progress and prospects. *International Journal of Urban Sciences*, 18:31–60, 09 2013.

Paulo Ribeiro, Fernando Fonseca, and Tânia Meireles. Sustainable mobility patterns to university campuses: Evaluation and constraints. *Case Studies on Transport Policy*, 8(2):639–647, 2020. ISSN 2213-624X. doi: https://doi.org/10.1016/j.cstp.2020.02.005. URL https://www.sciencedirect.com/science/article/pii/S2213624X20300158.

Jeppe Rich. Large-scale spatial population synthesis for denmark. *European Transport Research Review*, 10, 06 2018. doi: 10.1186/s12544-018-0336-2.

Jeff John Roberts. The gdpr is in effect: Should u.s. companies be afraid? https://web.archive.org/web/20180528010135/http://fortune.com/2018/05/24/the-gdpr-is-in-effect-should-u-s-companies-be-afraid/, 2018.

Robert W. Robinson. *Counting labeled acyclic digraphs*, pages 239–273. Academic Press, New York, 1973.

Robert W Robinson. Counting unlabeled acyclic digraphs. In Charles H. C. Little, editor, *Combinatorial Mathematics V*, pages 28–43, Berlin, Heidelberg, 1977. Springer Berlin Heidelberg. ISBN 978-3-540-37020-8.

Oliva G. Cantú Ros and Pedro García Albertos. D5. 4 enhanced version of matsim: synthetic population module. *Innovative Policy Modelling and Governance Tools for Sustainable Post-Crisis Urban Development (INSIGHT)*, 2016.

Ismaïl Saadi, Ahmed Mustafa, Jacques Teller, and Mario Cools. Forecasting travel behavior using markov chains-based approaches. *Transportation Research Part C: Emerging Technologies*, 69:402–417, 08 2016.

Ismaïl Saadi, Bilal Farooq, Ahmed Mustafa, Jacques Teller, and Mario Cools. An efficient hierarchical model for multi-source information fusion. *Expert Systems with Applications*, 110, 06 2018.

Aurore Sallard, Miloš Balać, and Sebastian Hörl. An open data-driven approach for travel demand synthesis: an application to são paulo. *Regional Studies, Regional Science*, 8(1):371–386, 2021. doi: 10.1080/21681376.2021.1968941. URL https://doi.org/10.1080/21681376.2021.1968941.

Mauro Scanagatta, Antonio Salmerón, and Fabio Stella. A survey on bayesian network structure learning from data. *Progress in Artificial Intelligence*, 8(4):425–439, 2019.

Patrick Schirmer, Michael van Eggermond, and Kay Axhausen. The role of location in residential location choice models: A review of literature. *Journal of Transport and Land Use*, 7, 07 2014. doi: 10.5198/jtlu.v7i2.740.

Markus Schläpfer, Lei Dong, Kevin O'Keeffe, Paolo Santi, Michael Szell, Hadrien Salat, Samuel Anklesaria, Mohammad Vazifeh, Carlo Ratti, and Geoffrey West. The universal visitation law of human mobility. *Nature*, 593:522–527, 05 2021.

Glenn R Shafer and Prakash P Shenoy. Probability propagation. *Annals of mathematics and Artificial Intelligence*, 2:327–351, 1990.

Dong Shaw and Geon Cho. The critical-item, upper bounds, and a branch-and-bound algorithm for the tree knapsack problem. *Networks*, 31:205–216, 07 1998. doi: 10.1002/(SICI)1097-0037(199807)31:43.0.CO;2-H.

Zhihao Shen, Wan Du, Xi Zhao, and Jianhua Zou. Dmm: Fast map matching for cellular data. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, MobiCom '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370851. doi: 10.1145/3372224.3421461. URL https://doi.org/10.1145/3372224.3421461.

Reza Shokri. Quantifying and protecting location privacy. *it - Information Technology*, 57:257–263, 01 2015. doi: 10.1515/itit-2015-0024.

Filippo Simini, Marta C. Gonzalez, Amos Maritan, and Albert-Laszlo Barabasi. A universal model for mobility and migration patterns. *Nature*, 484:96–100, 02 2012. doi: 10.1038/nature10856.

Richard Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *The Annals of Mathematical Statistics*, 35(2):876–879, 1964. ISSN 00034851. URL http://www.jstor.org/stable/2238545.

Chaoming Song, Tal Koren, Pu Wang, and Albert-Laszlo Barabasi. Modelling the scaling properties of human mobility. *Nature Physics*, 6, 10 2010.

Heinz Spiess. A maximum likelihood model for estimating origin-destination matrices. *Transportation Research Part B: Methodological*, 21(5):395–412, 1987. ISSN 0191-2615. doi: https://doi.org/10.1016/0191-2615(87)90037-3. URL https://www.sciencedirect.com/science/article/pii/0191261587900373.

Tim Spurr, Robert Chapleau, and Daniel Piché. Use of subway smart card transactions for the discovery and partial correction of travel survey bias. *Transportation Research Record*, 2405(1):57–67, 2014. doi: 10.3141/2405-08. URL https://doi.org/10.3141/2405-08.

Maria Stefanouli and Serafeim Polyzos. Gravity vs radiation model: two approaches on commuting in greece. *Transportation Research Procedia*, 24:65–72, 2017. ISSN 2352-1465. doi: https://doi.org/10.1016/j.trpro.2017.05.069. URL https://www.sciencedirect.com/science/article/pii/S2352146517303502. 3rd Conference on Sustainable Urban Mobility, 3rd CSUM 2016, 26 – 27 May 2016, Volos, Greece.

Peter Stopher, Camden Fitzgerald, and Min Xu. Assessing the accuracy of the sydney household travel survey with gps. *Transportation*, 34:723–741, 02 2007. doi: 10.1007/s11116-007-9126-8.

Lijun Sun and Alexander Erath. A bayesian network approach for population synthesis. *Transportation Research Part C Emerging Technologies*, 61:49–62, 10 2015.

Lijun Sun, Alexander Erath, and Ming Cai. A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B: Methodological*, 114:199–212, 2018.

Steffi Pauli Susanti and Fazat Nur Azizah. Imputation of missing value using dynamic bayesian network for multivariate time series data. In *2017 International Conference on Data and Software Engineering (ICoDSE)*, pages 1–5. IEEE, 2017.

Charles Sutton and Andrew McCallum. Piecewise training for undirected models. In *Conference on Uncertainty in Artificial Intelligence*, 2005. URL https://api.semanticscholar.org/CorpusID:1549479.

Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):571–588, oct 2002a. ISSN 0218-4885. doi: 10.1142/S021848850200165X. URL https://doi.org/10.1142/S021848850200165X.

Latanya Sweeney. K-anonymity: A model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, oct 2002b. ISSN 0218-4885. doi: 10.1142/S0218488502001648. URL https://doi.org/10.1142/S0218488502001648.

Ahmad Tavassoli, Mahmoud Mesbah, and Mark Hickman. Application of smart card data in validating a large-scale multi-modal transit assignment model. *Public Transport*, 10:1–21, 2018.

The OpenDP Team. The opendp white paper. https://projects.iq.harvard.edu/files/opendp/files/opendp_white_paper_11may2020.pdf, 2020. "[Online; accessed 24-August-2023]".

Manolis Terrovitis, Nikos Mamoulis, and Panos Kalnis. Privacy-preserving anonymization of set-valued data. *Proc. VLDB Endow.*, 1(1):115–125, aug 2008. ISSN 2150-8097. doi: 10.14778/1453856.1453874. URL https://doi.org/10.14778/1453856.1453874.

Marc Teyssier and Daphne Koller. Ordering-based search: A simple and effective algorithm for learning bayesian networks. *CoRR*, abs/1207.1429, 2012. URL http://arxiv.org/abs/1207.1429.

Mikkel Thorhauge, Habtamu Tilahun Kassahun, Elisabetta Cherchi, and Sonja Haustein. Mobility needs, activity patterns and activity flexibility: How subjective and objective constraints influence mode choice. *Transportation Research Part A: Policy and Practice*, 139:255–272, 2020. ISSN 0965-8564. doi: https://doi.org/10.1016/j.tra.2020.06.016. URL https://www.sciencedirect.com/science/article/pii/S0965856420306352.

TLC. New york city taxi and limousine commission trip record data. https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page, 2019.

Martin Trépanier, Robert Chapleau, and Nicolas Tranchant. Individual trip destination estimation in a transit smart card automated fare collection system. *Journal of Intelligent Transportation Systems: Technology, Planning and Operations*, 11(1):1–14, 04 2007. doi: 10.1080/15472450601122256.

Panagiotis Tsoleridis, Charisma Choudhury, and Stephane Hess. Probabilistic choice set formation incorporating activity spaces into the context of mode and destination choice modelling. *Journal of Transport Geography*, 108:103567, 04 2023. doi: 10.1016/j.jtrangeo.2023.103567.

Zhen Tu, Kai Zhao, Fengli Xu, Yong Li, Li Su, and Depeng Jin. Beyond k-anonymity: Protect your trajectory from semantic attack. In *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9, 2017. doi: 10.1109/SAHCN.2017.7964921.

Sarah Underwood. Can you locate your location data? *Commun. ACM*, 62(9):19–21, aug 2019. ISSN 0001-0782. doi: 10.1145/3344291. URL https://doi.org/10.1145/3344291.

United Nations. World urbanization prospects: The 2018 revision. Technical report, United Nations, 2018. [Online; accessed 30-November-2023].

Milad Vahidi and Yousef Shafahi. Time-dependent estimation of origin-destination matrices using partial path data and link counts. *Transportation*, pages 1–38, 2023. doi: 10.1007/s11116-023-10412-1.

David J van der Merwe and David Jacobus. *The use of partitioning strategies in local access telecommunication network problems and other applications*. PhD thesis, Potchefstroom Campus of the North-West University, 2007.

Nanne J van der Zipp and Rudi Hamerslag. Improved kalman filtering approach for estimating origin-destination matrices for freeway corridors. *Transportation Research Record*, 1443:54–64, 10 1994.

Daniele Veneziano and Marta C. Gonzalez. Trip length distribution under multiplicative spatial models of supply and demand: Theory and sensitivity analysis. *Computing Research Repository - CORR*, 11 2010.

Sofiia Vernydub. Modeling mode choice behavior after the introduction of new railway connections to munich airport. Master's thesis, TUM School of Engineering and Design, 2020.

Basil J. Vitins, Alexander Erath, and Kay W. Axhausen. Integration of a capacity-constrained workplace choice model: Recent developments and applications with an agent-based simulation in singapore. *Transportation Research Record*, 2564(1):1–13, 2016. doi: 10.3141/2564-01. URL https://doi.org/10.3141/2564-01.

Khoa D Vo, William HK Lam, Anthony Chen, and Hu Shao. A household optimum utility approach for modeling joint activity-travel choices in congested road networks. *Transportation Research Part B: Methodological*, 134:93–125, 2020.

Khoa D. Vo, William H.K. Lam, and Zhi-Chun Li. A mixed-equilibrium model of individual and household activity–travel choices in multimodal transportation networks. *Transportation Research Part C: Emerging Technologies*, 131:103337, 2021. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2021.103337. URL https://www.sciencedirect.com/science/article/pii/S0968090X21003405.

Feilong Wang and Cynthia Chen. On data processing required to derive mobility patterns from passively-generated mobile phone data. *Transportation Research Part C: Emerging Technologies*, 87:58–74, 2018. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2017.12.003. URL https://www.sciencedirect.com/science/article/pii/S0968090X17303637.

Huayong Wang, Francesco Calabrese, Giusy Di Lorenzo, and Carlo Ratti. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In *13th International IEEE Conference on Intelligent Transportation Systems*, pages 318–323, 2010. doi: 10.1109/ITSC.2010.5625188.

Yiyuan Wang, Bumsoo Lee, and Andrew Greenlee. The role of smart growth in residential location choice: Heterogeneity of location preferences in the chicago region. *Journal of Planning Education and Research*, page 0739456X2110176, 05 2021. doi: 10.1177/0739456X211017652.

Zhenzhen Wang, Sylvia Y He, and Yee Leung. Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society*, 11:141–155, 2018.

Zi-jia Wang, Xiao-hong Li, and Feng Chen. Impact evaluation of a mass transit fare change on demand and revenue utilizing smart card data. *Transportation Research Part A: Policy and Practice*, 77:213–224, 2015.

Wai Kit Wong, Nikos Mamoulis, and David Wai Lok Cheung. Non-homogeneous generalization in privacy preserving data publishing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, page 747–758, New York, NY, USA, 2010. Association for Computing Machinery. ISBN 9781450300322. doi: 10.1145/1807167.1807248. URL https://doi.org/10.1145/1807167.1807248.

Wei Wu, Yue Wang, Joao Bartolo Gomes, Dang The Anh, Spiros Antonatos, Mingqiang Xue, Peng Yang, Ghim Eng Yap, Xiaoli Li, Shonali Krishnaswamy, James Decraene, and Amy Shi-Nash. Oscillation resolution for mobile phone cellular tower data to enable mobility modelling. In *2014 IEEE 15th International Conference on Mobile Data Management*, volume 1, pages 321–328, 2014. doi: 10.1109/MDM.2014.46.

Zheli Xiong, Defu Lian, Enhong Chen, Gang Chen, and Xiaomin Cheng. A deeplearning framework for dynamic estimation of origin-destination sequence, 2023.

Boyam Yameogo, Pierre-Olivier Vandanjon, Pascal Gastineau, and Pierre Hankach. Generating a two-layered synthetic population for french municipalities: Results and evaluation of four synthetic reconstruction methods. *Journal of Artificial Societies and Social Simulation*, 24, 01 2021.

Boyam Fabrice Yameogo, Pascal Gastineau, Pierre Hankach, and Pierre Olivier Vandanjon. Comparing Methods for Generating a Two-Layered Synthetic Population. *Transportation Research Record*, 2675(1):pp. 136–147, January 2020.

Xianfeng Yang, Yang Lu, and Wei Hao. Origin-destination estimation using probe vehicle trajectory and link counts. *Journal of Advanced Transportation*, 2017:1–18, 01 2017. doi: 10.1155/2017/4341532.

Yingkun Yang, Chen Xiong, Junfan Zhuo, and Ming Cai. Detecting home and work locations from mobile phone cellular signaling data. *Mobile Information Systems*, 2021:5546329, 2021. doi: 10.1155/2021/5546329. URL https://doi.org/10.1155/2021/5546329.

Yingxiang Yang, Carlos Herrera, Nathan Eagle, and Marta C. González. Limits of predictability in commuting flows in the absence of data for calibration. *Scientific Reports*, 4(1):5662, 2014. doi: 10.1038/srep05662. URL https://doi.org/10.1038/srep05662.

Peijun Ye, Bin Tian, Yisheng Lv, Qijie Li, and Fei-Yue Wang. On iterative proportional updating: Limitations and improvements for general population synthesis. *IEEE Transactions on Cybernetics*, PP:1–10, 06 2020. doi: 10.1109/TCYB.2020.2991427.

Xin Ye, Karthik Konduri, Ram Pendyala, Bhargava Sana, and Paul Waddell. Methodology to match distributions of both household and person attributes in generation of synthetic populations. In *Transportation Research Board 88th Annual Meeting*, 2009.

Ling Yin, Qian Wang, Shih-Lung Shaw, Zhixiang Fang, Jinxing Hu, Ye Tao, and Wei Wang. Re-identification risk versus data utility for aggregated mobility research using mobile phone location data. *PLOS ONE*, 10 (10):1–23, 10 2015. doi: 10.1371/journal.pone.0140589. URL https://doi.org/10.1371/journal.pone.0140589.

Mogeng Yin, Madeleine Sheehan, Sid Feygin, Jean-Francois Paiement, and Alexei Pozdnoukhov. A generative model of urban activities from cellular data. *IEEE Transactions on Intelligent Transportation Systems*, PP: 1–15, 05 2017. doi: 10.1109/TITS.2017.2695438.

Mogeng Yin, Madeleine Sheehan, Sidney Feygin, Jean-François Paiement, and Alexei Pozdnoukhov. A generative model of urban activities from cellular data. *IEEE Transactions on Intelligent Transportation Systems*, 19(6):1682–1696, 2018. doi: 10.1109/TITS.2017.2695438.

Seo Youn Yoon, Kathleen Deutsch, Yali Chen, and Konstadinos G. Goulias. Feasibility of using time–space prism to represent available opportunities and choice sets for destination choice models in the context of dynamic urban environments. *Transportation*, 39(4):807–823, 2012.

Hui Zang and Jean Bolot. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the Annual International Conference on Mobile Computing and Networking, MOBICOM*, pages 145–156, 09 2011. doi: 10.1145/2030613.2030630.

Danqing Zhang, Junyu Cao, Sid Feygin, Dounan Tang, Max Shen, and Alexei Pozdnoukhov. Connected population synthesis for transportation simulation. *SSRN Electronic Journal*, 01 2019.

Yang Zhang, Dongchao Ma, Fan Zhang, Yan Li, Yuzhu Jin, Lihua Song, and Laizhong Cui. A research for travel mode identification based on cellular signaling data. In *2022 IEEE Smartworld, Ubiquitous Intelligence & Computing, Scalable Computing & Communications, Digital Twin, Privacy Computing, Metaverse, Autonomous & Trusted Vehicles (SmartWorld/UIC/ScalCom/DigitalTwin/PriComp/Meta)*, pages 318–325, 2022. doi: 10.1109/SmartWorld-UIC-ATC-ScalCom-DigitalTwin-PriComp-Metaverse56740.2022.00067.

Peng Zhuang, Yi Shang, and Bei Hua. Statistical methods to estimate vehicle count using traffic cameras. *Multidimensional Systems and Signal Processing*, 20:121–133, 2009.

Michael Zilske and Kai Nagel. A simulation-based approach for constructing all-day travel chains from mobile phone data. *Procedia Computer Science*, 52:468–475, 2015.

Chapter 7

Appendices

7.1 Optimization and duality

In this section, we describe the basics of optimization and the general context of duality, which are a key component of our anonymisation approach. We will focus on the simple case of a minimisation problem with k inequality constraints:

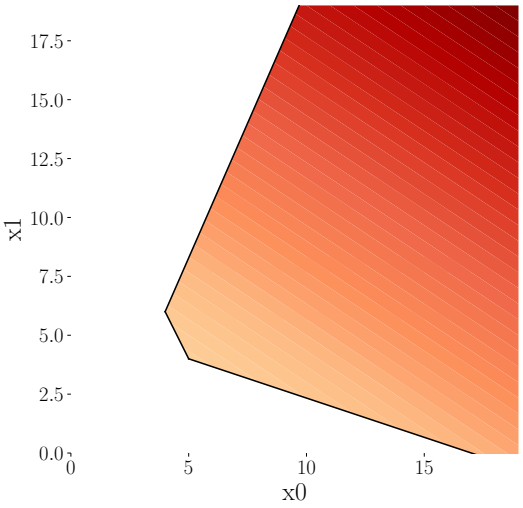
$$\begin{aligned} &\min_x f_0(x) \\ &s.t. \forall i = 1, \dots, k \, f_i(x) \leq 0 \end{aligned} \tag{7.1}$$

where $x \in \mathbb{R}^n$ is a variable in any dimension, and the functions $f_i : \mathbb{R}^n \to \mathbb{R}$, are general functions. In optimization terminology, we say that problem 7.1 finds the minimum of the **objective function** f_0. Any x satisfying the constraint $\forall i = 1, \dots, k, f_i(x) \leq 0$ is called **feasible**, and an x solving the problem is called **optimal**. The value of the objective function $f_0(x^*)$ for an optimal solution x^* is called the **value** of problem 7.1.

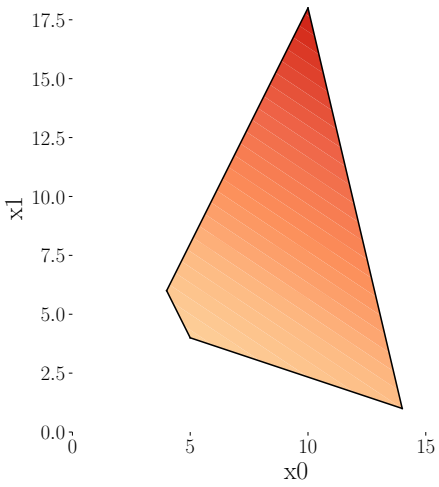
7.1.1 Solving a primal problem

Problem 7.1 is a very general problem and depending on the properties of f, various approaches can be used to explore the **feasible set** described by $\{x, \forall i = 1, \dots, k, f_i(x) \leq 0\}$. The simplest form of this problem is when all function f_i are linear, in which case we call the problem a linear program. For each $1 \leq i \leq k$, the set of x satisfying the constraint $f_i(x) \leq 0$ is then a half-space in $\mathbb{R}^n$. The intersection of all these half-spaces forms the feasible set. The gradient of the objective function f_0 represents a direction to take. The graphic representation of the situation (Fig. 7.1) gives an intuitive result that the problem is either not bounded, *i.e.,* x can go in the direction of the gradient indefinitely without encountering any barrier, or the problem is bounded and at least one solution lies on a vertex of the feasible set.

This has inspired the first linear programming algorithm, the simplex algorithm (Dantzig, 1990), which visits the vertices of the feasible set until finding the optimal one. Later, the ellipsoid method (Khachiyan, 1980) has been developed that is much less efficient than the



(a) Unbounded problem: given a feasible solution, it is always possible to follow the gradient and find a better solution.



(b) Bounded problem: the optimal solution exists and is found on at least one vertex of the boundary.

Figure 7.1: Examples of an unbounded and a bounded linear programs. Constraints are represented as lines where it is allowed to be on one side only. The intersection of the constraints is a polygon in which the objective function is represented as a color gradient.

simplex algorithm but is proven to terminate in polynomial time in the worst case. The latest approach to solve linear programs is the interior-point method, which also runs in polynomial time but is also efficient in practice (Karmarkar, 1990). Interior point methods also apply when f_0 is not necessarily linear but still convex. Note that in these cases the difficulty comes not from the objective function f_0, which is very simple, but from the constraints $f_{i\geq 1}$ that describe a feasible set in high dimension such that it is not trivial to find a feasible x in the first place.

Another approach for convex objective functions is the **gradient descent** algorithm. This method arises from the intuitive remark that the minimum of a function can be found if we keep going down. The steps are as follows:

1. Choose a step size $\lambda > 0$
2. Choose a starting point $x_0 \in \mathbb{R}^n$
3. Compute the gradient $g = \nabla f_0(x_i)$
4. Update $x_{i+1} \leftarrow x_i - \lambda g$
5. Iterate from step 3 until convergence.

In the case of a convex differentiable function f_0, the convergence is guaranteed and characterized by a gradient $\nabla f_0(x_i) = 0$. In the practical case, it suffice to stop when the gradient is under a small threshold. The gradient descent algorithm can be adapted as a constrained version for a convex feasible set, where at each step the new point x_{i+1} is projected on the convex set. The projected gradient descent algorithm is guaranteed to find the minimum of the convex function among the feasible set if this set is convex. However, the main usage

of gradient descent in modern optimization problems is the learning of deep neural networks, where the error function is not guaranteed to be convex and is even known to feature numerous local minima. Numerous heuristic adaptation exist to handle these cases, the most prominent as of today being Adaptive Moment Estimation (ADAM) (Kingma and Ba, 2014).

In the case where nothing is known about f_0, not even if it is derivable, problem 7.1 becomes a "blackbox" problem. Such a problem requires **derivative-free** optimization methods, which explore the feasible set starting from a solution and generating neighboring solutions using a set of rules. Heuristic rules are required to choose the direction to explore among all the possible neighbors at each step. In the absence of a gradient to provide a stop condition such as $\nabla f_0 = 0$, the optimal solution is selected when the algorithm either stops finding other solutions than the current one, or when it cannot find a neighbor that has a better value than the current solution. Among the variety of such approaches, we can mention some broad families (Larson et al., 2019):

- **Random search algorithms**, which generate a new neighboring solution at random until one of them has a better value, and then iterate from this new solution.
- **Tabu search algorithms**, which keep track of the explored possibilities to avoid visiting them again.
- **Simulated annealing**, which performs a random search with a probability to accept the new solution even if the value is not better than the current one. The acceptance probability decreases at each step in order to ensure convergence to a solution.
- **Swarm algorithms**, which implement several random search at the same time with a possibility for each instance to depend on the result of the others to decide the next step.
- **Genetic algorithms**, in which a population of solutions is generated and a subset of the solutions with the best values is select to "reproduce", *i.e.,* the next population is created from variations of those best-fit solutions.

As these approaches are computationally very expensive, there are recommended only in situations where the objective function f_0 is not differentiable. An example of such a situation would be the learning of the structure of probabilistic graphical model, where the variable x is in $\{0,1\}^n$ indicating for each possible edge if it must be drawn or not. As there is no way of knowing if the performance of the model will increase or decrease when adding an edge to the graph, structure learning relies on blackbox methods to explore the possible structures. In this situation, the operations defining neighboring solutions are generally adding, removing, or inverting an edge.

7.1.2 Relaxation and dual problem

Relaxing the constraints: As it has been mentioned previously, the main complexity of an optimization problem such as problem 7.1 resides in the constraints (Lemaréchal, 2001). The feasibility of a solution x is a discrete state that carries no information on "how feasible" or "how unfeasible" x is, which could help us find the set when we are outside of it. This can make exploring the feasible set a hard problem in itself, whereas evaluating the objective function is comparatively easy.

When a constraint is too hard to evaluate, a common approach to solve the problem is to relax it (Geoffrion, 2009). The Lagrangian relaxation of a constraint consists in adding a cost $\lambda \geq 0$ to the objective function that is positive if the constraint is not satisfied, thus penalizing solutions that are not feasible in the original formulation of the problem. The relaxation of all constraints from problem 7.1 gives problem 7.2:

$$\inf_x f_0(x) + \sum_{i=1}^{k} \lambda_i f_i(x). \tag{7.2}$$

Instead of minimizing the objective function f_0, the relaxed problem fixes some penalization factors $\lambda_1, \ldots, \lambda_k \geq 0$ and aims to find the x that minimize the **Lagrangian function** L:

$$L : x, \lambda_1, \ldots, \lambda_k \mapsto f_0(x) + \sum_{i=1}^{k} \lambda_i f_i(x).$$

Relaxing the constraints can be very powerful when the functions f_i are simple in the first place. For example in the case of linear programming where all f_i are linear, the infimum $\inf_x L(x, \lambda_1, \ldots, \lambda_k)$ can be expressed as a closed-form, linear expression. Note that in that case, it is $-\infty$ unless the λ_i are carefully chosen so that the slope of $L(x)$ is null.

Dual problem: Recall that if x is feasible for problem 7.1, then all $f_i(x)$ are negative. This in turn implies that $L(x, \lambda_1, \ldots, \lambda_k) \leq f_0(x)$ for any value of the λ_i. If we note $G(\lambda) = \inf_x L(x, \lambda_1, \ldots, \lambda_k)$ the solution to the relaxed problem 7.2, it follows by mean of the infimum bound that $G(\lambda)$ is necessarily lower than the value of the objective function $f_0(x)$ for any feasible x:

$$\forall x \text{ feasible}, \forall \lambda_1 \ldots \lambda_k \geq 0, G(\lambda_1 \ldots \lambda_k) \leq f_0(x).$$

And in particular for an optimal solution x^* to problem 7.1, which verifies $\forall x$ feasible, $f_0(x)\ leq f_0(x^*)$:

$$\forall \lambda_1 \ldots \lambda_k \geq 0, G(\lambda_1 \ldots \lambda_k) \leq f_0(x^*).$$

It can then be of interest to find the values of λ that maximize $G(\lambda_1 \ldots \lambda_k)$. This third optimization problem is called the **dual problem of 7.1** and is of the form:

$$\begin{aligned} &\max_{\lambda_1, \ldots, \lambda_k} \inf_x f_0(x) + \sum_{i=1}^{k} \lambda_i f_i(x) \\ &s.t. \forall i = 1, \ldots, k, \lambda_i \geq 0 \end{aligned} \tag{7.3}$$

In the worst case, solving the dual problem 7.3 can give us a lower bound on the minimum value we can expect from f_0. In the best case, called **strong duality**, the maximum value of G coincides with the minimum value of f_0. In any case, solving problem 7.3 is generally easier than solving the primal problem 7.1. Indeed, the constraints of problem 7.1 can be numerous and convoluted, whereas there are only n constraints in problem 7.3 (one for each variable in the primal problem) and they are linear.

Feasibility of the dual solution: We could be tempted to solve the dual problem in order to obtain a solution for the primal. Indeed, if the infimum is attained, the solution λ^* implies that there exist a x^d such that $G(\lambda^*) = L(x^d, \lambda^*)$. We could expect that this solution x^d is a feasible solution of the primal problem, but this is actually not necessarily the case. In the following, we summarize what we know about x^d.

If x^* is the optimal solution of the primal problem, we know that $L(x^d, \lambda^*) \leq L(x^*, \lambda^*)$, which translates to:

$$f_0(x^d) + \sum_{i=1}^{k} \lambda_i^* f_i(x^d) \leq f_0(x^*) + \sum_{i=1}^{k} \lambda_i^* f_i(x^*). \tag{7.4}$$

This gives us an upper bound for the sum of penalties associated to x^d:

$$\sum_{i=1}^{k} \lambda_i^* f_i(x^d) \leq f_0(x^*) - f_0(x^d) + \sum_{i=1}^{k} \lambda_i^* f_i(x^*) \tag{7.5}$$

where x^* is feasible so $\sum_{i=1}^{k} \lambda_i^* f_i(x^*) \leq 0$, and x^* is the optimal solution of the primal so $f_0(x^*) \leq f_0(x^d)$. It follows:

$$\sum_{i=1}^{k} \lambda_i^* f_i(x^d) \leq 0. \tag{7.6}$$

The sum of penalties being negative is a necessary condition for the solution x^d to be feasible but it is not sufficient, as the feasibility is characterised by having all f_i negative. In the case where there is only one constraint, Eq. 7.6 tells us that the solution x^d implied by the dual is feasible, without necessarily being the optimal solution.

Even if strong duality holds, Eq. 7.6 only translates to:

$$\sum_{i=1}^{k} \lambda_i^* f_i(x^d) = f_0(x^*) - f_0(x^d), \tag{7.7}$$

which means that depending of the profiles of the constraints and the objective functions, they could balance out at other points than the optimal primal solution x^* and the solution x^d could be different from x^* in such a way that it is neither optimal nor feasible. In conclusion, although the dual problem can be easier to solve than the primal, the solution x^d implied by the solving is not necessarily usable in the general case. In practice, x^d can still be a reasonably good solution whose main advantage is the low computational cost, but it requires additional checks to ensure that it is feasible and needs to be ensured on a case-by-case basis.

7.1.3 Formulations of primal and dual problems

An important aspect of optimization problems is that they can be formulated in many different, equivalent forms. This has several convenient implications:

- First, numerous applications will give a problem that does not look exactly like the one presented in this chapter. By reformulating it, we can reduce a problem to a known form, ensuring that the properties that are discussed here are also applicable to the

more complex forms. Reducing a problem to a more familiar form for which a specific efficient algorithm exist is also a way of solving it.

- Second, the computational cost of solving a problem depends greatly on the formulation. Reformulating the problem in another form can speed-up its solving by standard solvers.
- Third, the dual formulation of a problem depends on the formulation of its primal. By reformulating the primal problem, the associated dual can itself be reduced to a known problem or be made cheaper to solve.

Here are some examples of the possible transformation that are applicable to an optimization problem: The first obvious possible transformation is that a problem with a constraint of the form $f_i(x) \geq 0$ is equivalent to the same problem with the constraint $-f_i(x) \leq 0$ instead. An equality constraint $f_i(x) = 0$ can be expressed as two inequality constraints $f_i(x) \leq 0$ and $f_i(x) \geq 0$, or we can introduce a slack variable y that does not appear in the objective function but change the constraint to $f_i(x) + y = 0$ and $y \geq 0$.

In general, it is always possible to add intermediary variables in the formulation of the objective function, with an equality constraint specifying the value of this additional variable with respect to to initial ones. The opposite is also possible, and it can be valuable to remove a variable that is subject to an equality constraint, as it means that it can be replaced by its expression in the formulation of the objective function. This allows to reduce the complexity of the problem by one variable and one constraint.

In some cases such a maximum likelihood problems, the objective function f_0 contains a lot of products that make it expensive to compute and hard to differentiate. It is then common to solve the problem $\min_x \log(f_0(x))$ instead, which accepts the same solution x^*. More generally, given any strictly increasing function g, the optimal solution $x^* = \min_x f_0(x)$ is also the optimal solution of the problem $x^* = \min_x g(f_0(x))$. Likewise, if g is strictly decreasing, then the optimal solution $x^* = \min_x f_0(x)$ is the optimal solution of the maximization problem $x^* = \max_x g(f_0(x))$. Conversely, applying a strictly decreasing function to a maximization problem transforms it into a minimization problem.

More advanced transformations are also possible. For example, although the absolute value function is convex, it can be useful to get rid of expressions such as $|x|$ in the formulation of the objective function, especially if doing so can give way to a completely linear program. A common trick to achieve this in the case of minimization problems is to replace the variable x by $x = x^+ - x^-$ where $x^+ \geq 0$ and $x^- \geq 0$ represent the positive and negative parts of x, respectively. We can then replace x by $x = x^+ + x^-$. Solving this modified problem will yield the same optimal solution in the sense that either x^{*+} or x^{*-} will be equal to 0: if the minimum of x^{*+} and x^{*-} is $\delta > 0$, then considering $x'^{+} = x^{*+} - \delta$ and $x'^{-} = x^{*-} - \delta$ gives the same solution $x' = x'^{+} - x'^{-} = x^*$ but the objective function is reduced by 2δ, so x^* cannot be optimal. Note that this trick works only for minimization problems where the objective function is increasing with respect to $|x|$.

As for the dual problem, it is not necessary to relax all the constraints. By relaxing only a subset of the constraints of a problem, it is possible to formulate a great variety of different dual problems associated to a same primal. It is then more adequate to speak of *a* dual than of *the* dual of a problem. Depending on the context, one of these problem may be easier to

solve than the others, or be reduced to another, well-known problem. However, there is no readily applicable method to get the best dual problem in the general case. Care must always be given to of the way the problems are formulated and which constraints are relaxed for the dual. In chapter 3, we will see an example of such a reformulation in which the problem at hand is reformulated as a variation of the knapsack problem (the precedence constraint knapsack problem), and how a particular algorithm (the single breakpoint algorithm) solves it dual while implying a feasible solution for the primal.

7.2 Probabilistic graph models

Another tool used in this work are probabilistic graph models to represent the dependencies between variables. A rich literature in travel demand considers the data at hand as the realization of an intricate probability law that needs to be approximated (Sun and Erath, 2015; Joubert and De Waal, 2020; Anda et al., 2021). This data can be the socio-economic variables describing a population of individuals, of the times, transport modes, purposes, or locations of their activities. As noted in Sec. 2.3, approximating the law governing this data allows to understand its structure and eventually to generate more of it, which is extremely valuable if the task is to predict the behavior of the system in an hypothetical situation. Drawing a graph where each variable is a node and each edge indicates that a variable is dependent on the other is a first step in illustrating such an intricate joint distribution. This type of graph is called a **Bayesian network**, and it can look like the illustration in Fig. 7.2. This graph represents the joint distribution of ten variables, where the conditional distribution of one variable with respect to the others is actually equal to the conditional distribution with respect to its parents in the graph. For example, here $p(X_{10}|X_1, \ldots, X_9) = p(X_{10}|X_7, X_9)$. In the case of discrete variables, we represent these conditional distributions by **conditional probability tables**.

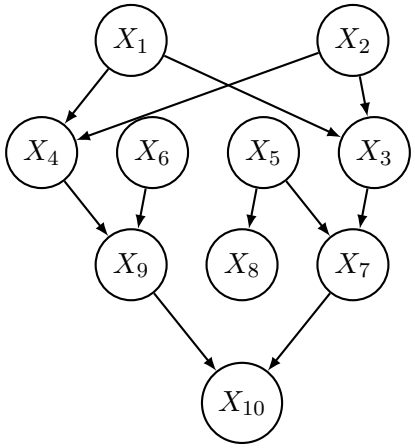


Figure 7.2: Example of Bayesian network.

7.2.1 Bayesian networks

Bayesian networks are not really a way of representing the joint distribution of a data sample. They are rather a way of representing the most important links, while actively discarding the others. Indeed, if we measure the quality of a Bayesian network by its capacity to

approximate a given probability distribution in its details, then a Bayesian network with more edges performs necessarily better than the same network with fewer edges, as it incorporates more information on the sample we have at hand. In the same way as in classical machine learning, adding such non-essential edges a form of over-fitting. It is also a loss for the explainability power of the Bayesian network, of which a great strength is to allow to explain *why* a variable has a certain value (Pearl, 2009). It is worth noting that in Bayesian networks, the absence of an edge is as informative as its presence. Indeed, it is easy to build a conditional probability table such that the explanatory variable does not actually influence the outcome, meaning that in a Bayesian network that is not carefully built, some edges can appear that do not actually exist. In contrast, the absence of edge represents the strong assumption of independence between two variables, which is very informative in itself. More generally, Pearl (1988) introduces the concept of Markov blanket $\mathcal{B}_v$ of a node v to refer to a set of nodes that gives the maximal information on node v: given the values of all the nodes in $\mathcal{B}_v$, the distribution of v does not depend on any other variables. The minimal Markov blanket of a node v is the set containing its parents, its children, and the other parents of its children, as illustrated in Fig. 7.3.

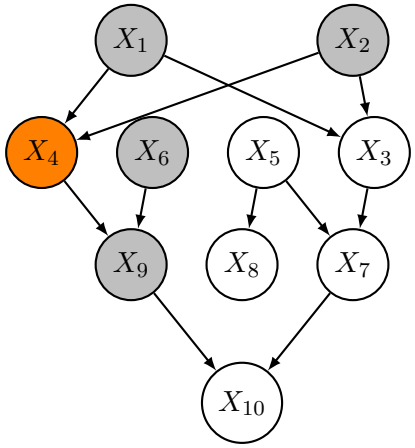


Figure 7.3: Markov blanket of node X_4

Sampling Bayesian networks: Bayesian networks can be used to sample new data from the approximated distribution by simply drawing the variables in order. This is always possible as they are required to be acyclic, so there exist at least one variable that does not depend on anything else, and for any variable that depend on some parents, by following back the chain of dependency we necessarily find parents that do not depend on any variable. Bayesian networks can also be used to generate missing data in an incomplete observation (Susanti and Azizah, 2017). This is much less straightforward, and the best way to do this is with the sum-product algorithm, as would be the case for a Markov chain for example (Norris, 1998). The sum-product algorithm only works for tree-like graphs, although it can be used in general graphs to yield an approximate solution in an algorithm called loopy belief propagation (Jordan, 2015). For exact inference on a Bayesian network, we need to represent it as a tree. More formally, a Bayesian network can be represented as a junction tree (Lauritzen and Spiegelhalter, 1988). A junction tree $T = (V_T, E_T)$ for a graph $G = (V_G, E_G)$ is a tree whose nodes represent subsets of nodes of G with the following properties:

- Tree: for any two nodes $A, B \in V_T$, there exists a unique path from A to B, noted path$(A \to B)$
- Running intersection: for any two nodes $A, B \in V_T$, for all $C \in \text{path}(A \to B)$, $A \cap B \subset C$
- Covering: for any node $v \in V_G$, there exists at least one node $A \in V_T$ such that $\text{fa}(v) \subset A$, where $\text{fa}(v)$ is the **family** of v, *i.e.* the set of v and its parents.

An example is given in 7.4. A junction tree can be intuitively seen as a tree T in which each node $v \in V$ is represented by a subtree T_v of T, and v_1, v_2 are adjacent in G implies that T_{v_1} and T_{v_2} intersect in T. A trivial junction tree for any graph G is a single node $C = V$. However this particular junction tree will not be useful, and we usually look for a junction tree whose largest node C has minimal cardinal. Constructing a relevant junction tree is a problem addressed by the Shafer-Shenoy algorithm (Shafer and Shenoy, 1990) or the Hugin algorithm (Jensen et al., 1994).

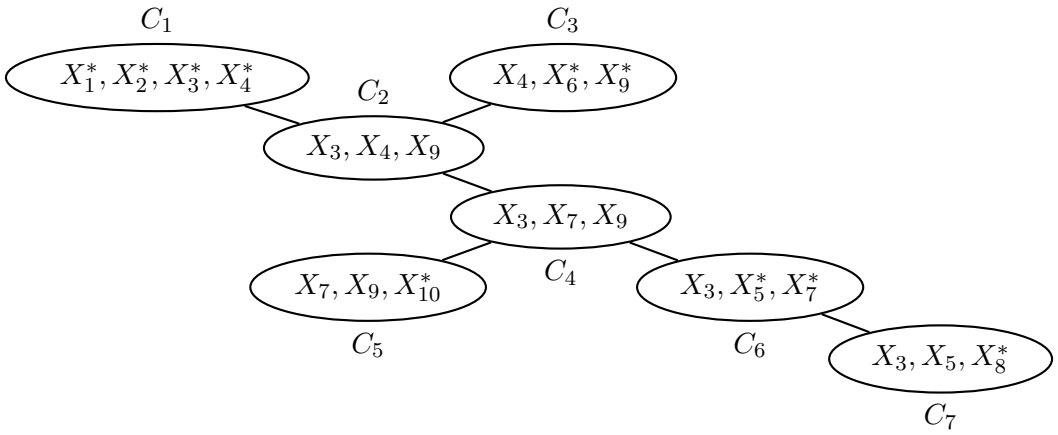


Figure 7.4: A possible junction tree for the Bayesian network illustrated in Fig. 7.2. Example taken from Nuel (2012).

Message propagation algorithms such as the sum-product algorithm allow to estimate the complete probability law of one or more nodes of the network, which we can then draw from. But if the task is only to draw a sample, the true joint probability is not required. It is enough to know the conditional distributions of each variable with respect to the others, which are known since they are the conditional probability tables defining the network. The approaches used in this cases are known as **Markov Chain Monte Carlo (MCMC)** methods. In this work, we are mostly interested in the **Gibbs sampling** algorithm (Geman and Geman, 1984):

1. Start with random values on the known that need to be sampled and fixed values on the nodes that are known.
2. Select a variable v among the ones that needs to be sampled.
3. Draw a new value for this variable conditioned on all the other variables.
4. Repeat at step 2.

Gibbs sampling builds a Markov chain where each state contains value attribution for all the variables. It can be proven that the stationary distribution of this Markov chain is the joint

distribution f of the variable: Consider two states x and y of the Markov chain. We note $y \sim_i x$ if x and y are equal for all variables except in variable i. Note that $\sim_i$ is an equivalence relation, *i.e.,* that in particular that it is transitive and symmetric. We note p_{xy} the probability of a node of the Markov chain being set to y knowing that the previous node is set to x. That is:

$$p_{xy} = \frac{1}{K} \frac{f(y)}{\sum_{z, z \sim_i x} f(z)}, \tag{7.8}$$

where K is a normalization constant. To prove that f is the stationary distribution of the Markov chain, we prove that $f(x)p_{xy} = f(y)p_{yx}$ (Lindley, 2004). This is true as:

$$f(x)p_{xy} = \frac{1}{K} \frac{f(x)f(y)}{\sum_{z, z \sim_i x} f(z)} = \frac{1}{K} \frac{f(y)f(x)}{\sum_{z, z \sim_i y} f(z)} = f(y)p_{yx}.$$

We proved that the distribution of the states of the Markov chain converges to the joint distribution of the variables. This means that to get a sample that follows the joint distribution represented by the Bayesian network, we can perform Gibbs sampling and just take the state of the network at any point after a large number of step. This large number of steps required to get from the initial random distribution to the stationary distribution is called the warm-up phase (Carter and Kohn, 1994). In practice, if we want several samples from the same distribution, it is recommended to also wait for a number of steps before taking another state after the first, to reduce the obvious correlations that occur between closed nodes of the Markov chain (Bishop and Nasrabadi, 2006).

Learning Bayesian networks: Bayesian networks are a great way of combining expert knowledge with the available data, especially if it is incomplete: an educated guess or prior knowledge on the mechanism we want to model can give the structure of the network, and then the implied conditional probability tables can be estimated by simply counting the number of occurrences in the data. This very straightforward estimator is conveniently the maximum likelihood estimator. Note that this estimation can make use of an incomplete row in the data, since the conditional probability table of a given variable can take into account any row that has an observation for the variable and all its parents. When the available data is severely incomplete, or features a latent variable that is never observed (either due to a technical problem in the data collection or as an hypothesis in the conception of the graph), then the conditionals cannot all be estimated this way, and we must rely on the Expectation-Maximisation (EM) Algorithm. The EM Algorithm is relatively easy to interpret for Bayesian networks: the E step simply consists in completing the missing observations *via* belief propagation, and the M step estimates the parameters as if the data was complete (Nuel, 2012).

The structure of the Bayesian network can also be learned, although for this task complete data is required. It is a much more difficult task due to the combinatorial number of structures that could be explored.

As we noted above, Bayesian networks can only yield better predictions when including more edges. Because of this, when learning the structure of a Bayesian network we refer to old-school criteria that penalize the number of edges of the model as well as encourage its accuracy. The two most famous of those criteria, developed with the mindset that the simplicity of a model is the only thing preventing it from overfitting, are the **Bayesian Information**

Criterion (Neath and Cavanaugh, 2012) and the **Akaike Information Criterion** (Bozdogan, 1987), best known as **BIC** and **AIC**. They both reward the likelihood of the model while directly penalizing the number of parameters:

$$AIC = 2p - 2\ln(L) \tag{7.9}$$

$$BIC = p\ln(n) - 2\ln(L). \tag{7.10}$$

Where p is the number of parameters, n is the number of rows in the observed data x, and $L = p(x|\theta)$ is the likelihood of the model given the observed data. In the case of a Bayesian network with only discrete variables, the number of parameters p is the number of cells in the conditional probability tables that are implied by the graph. The most prominent approach to structure learning is to use a variant of Tabu search with a wide variety of tricks (Teyssier and Koller, 2012) to explore the solution space with maximum efficiency (Sun and Erath, 2015; Joubert and De Waal, 2020). Zhang et al. (2019) points out that structural learning is seemingly not robust when subject to bootstrapping. However, it is important to note that some very different structures of Bayesian Network may yield similar joint distributions. It is then not obvious to assess the true robustness of a structure learning algorithm.

7.2.2 Undirected graphical models

Bayesian networks are required to be oriented and acyclic. The orientation of the edges is important as it gives the structure of the conditional probability table by showing which is the explained and which are the explanatory variables, and when generating data from scratch it shows with which nodes we can start the drawing. It is also possible to represent a distribution with a non-oriented graph, in which case an edge still represents a dependency link but the idea of conditional probability table is lost. We could think of interpreting an undirected graphical model as having one conditional probability law for each node, which depends on its neighbors, but it would be hard to ensure that such conditionals are consistent with each other and that their product ultimately forms a probability law. We can retain the core property that the joint distribution is obtained by the product of all the functions if we drop the constraint that those functions need to sum to one. In that case, we call them potentials, and the joint distribution is obtained by the product of all the potentials, up to a normalisation factor. The potentials in an undirected graph are not associated to the nodes, as would be the case in a directed graph. An edge between two node rather indicates that there should be a factor depending on these two nodes, and likewise a group of totally interconnected nodes indicates that there should be a factor containing all those nodes. The potentials are thus associated to the cliques of the graph, preferably the maximum cliques (those who are not a subpart of a bigger clique). Following the remark on directed graphs that a link can actually be parameterized so that the linked variables are actually independent, in directed graphs, a clique can actually represent several structures of potentials. For example, both the factorization $p(x_1, x_2, x_3) = \frac{1}{Z}\phi(x_1, x_2, x_3)$ and its more restrictive form $p(x_1, x_2, x_3) = \frac{1}{Z}\phi_1(x_1, x_2) \times \phi_2(x_2, x_3) \times \phi_3(x_1, x_3)$ are represented by the graph where the three variables X_1, X_2 and X_3 form a clique, as illustrated in Fig. 7.5.

For this reason, it is more informative to explicitly represent the potentials on the graph. Doing so gives a **factor graph**, a bipartite graph where the nodes representing the variables are connected to nodes representing the potentials, or factors. Each factor is a function depending

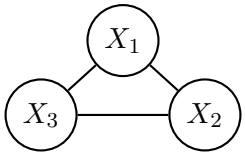
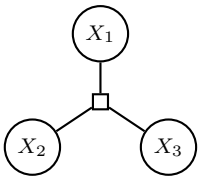
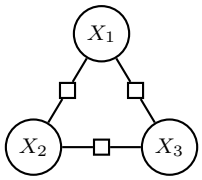


Figure 7.5: An undirected graph that can actually represent several possible factorizations.

on the variables it is connected to in the graph, and the joint distribution of the variables is once again given by the product of the factors, to a normalization. The two factorizations of the example are represented by two distinct factor graphs, one with only one factor and the other with three (Fig. 7.6).



(a) No factorization: $p(x_1, x_2, x_3) = \frac{1}{Z}\phi(x_1, x_2, x_3)$



(b) Full factorization: $p(x_1, x_2, x_3) = \frac{1}{Z}\phi_1(x_1, x_2) \times \phi_2(x_2, x_3) \times \phi_3(x_1, x_3)$

Figure 7.6: Two factorizations that are represented by the same undirected graph but by distinct factor graphs.

Probabilistic graph models have been used intensively for the generation of synthetic data, whether for the synthetic population in itself with its socio-economical variables, the activity chains, or the locations of the activities.

7.2.3 Limitations

Using graph models in practice is limited by the scalability of the structure learning. Even when limiting ourselves to bayesian networks, the number of possible structures grows super-exponentially with the number of nodes. According to the enumeration found by Robinson (1973, 1977), there are 4×10^{18} possible structures for 10 nodes, and more than 10^{100} structures for 24 nodes. Apart from the previously mentionned tabu search, searching this feasible set efficiently is the subject of active research, but the most scalable approaches can be applied for problems of only up to 200 nodes (Scanagatta et al., 2019). In the case of incomplete data, the best approaches can scale up to 40 nodes. In this section, we focused on discrete networks where each node is a multinomial variable and the dependencies are implemented by conditional probability tables. Most results are applicable only in this situation, and the usual way of handling continuous variable is to discretize them (Beuzen et al., 2018). Structure learning with discretization of continuous variable has been explored by Fernández et al. (2013). When the discretization is deemed unnacceptable, some approaches such as Cobb et al. (2007) aim at developping mixed networks.

7.3 Detailing of the anonymization results

We detail the results of each solver over each dataset in table 7.1.

7.4 Temporal analysis is still possible even though separate OD-matrices are generalised differently for each timestep

In this section, we illustrate how separating the flows into distinct OD-matrices corresponding to time steps does not stop us from doing temporal analysis. Indeed, the reconstruction process described in Section 3.6 allows us to retrieve an estimation of the flows between any arbitrary couple of zones, which we can choose to be consistent across timesteps. By reconstructing anonymised OD flows on the initial zoning, we can compare them with the initial flows. An exemple of this is given in Fig. 7.7, which represents the mean volume of some example flows throughout the week, as obtained from the original data and from the data anonymised with ATG-dual.

7.5 Tweaking of the probability law given by the factor graphs

In this section, we detail the adaptation applied to our factor graph approach based on the remark that the structure of the model yields probability laws that are inherently of the wrong dimension: The marginal probabilities for one given state are proportional to the product of several factors while they should be only proportional to one of them. We first illustrate the problem in a situation where all individuals share the same agenda and we then generalize to a population where agents have distinct agendas.

Population with a single agenda: We consider a population of individuals who all share the same agenda «h-w-h-w-o-h»: Starting from home activity `h`, the agents go to work `w`, then back home, back to work, then to an "other" activity `o`, and finally back home. By modeling the agenda as a list of activities that share locations, we obtain six variables, some of which are required to be equal (in this paper, we say that they are linked). The corresponding factor graph is given in Fig. 7.8. We number the factors by the states they are connected to, which is natural here as two factors connected to the same states will have equal estimators from our data: for example the factor $f_{0,1}$ is a function taking two values and returning a potential for the couple (Z_0, Z_1) having these values. We are interested in the probability law of the work location given a home location h and a secondary location s. The joint probability law $\hat{p}$ is proportional to the product of all the factors, so by conditioning, the marginal law of (Z_1, Z_3) is proportional to the factors depending on Z_1 or Z_3:

$$\begin{aligned}\hat{p}(Z_1 = Z_3 = w | Z_0 = Z_2 = h, Z_4 = s) \propto & f_{0,1}(h,w) \times f_{1,2}(w,h) \times f_{1,2}(w,h) \\ & \times f_{2,3}(h,w) \times f_{3,4}(w,s).\end{aligned} \tag{7.11}$$

The factors are estimated by the flows observed from OD matrices:

$$f_{i,j}(o,d) = v_{t_{ij}}(o \to d), \tag{7.12}$$

where t_{ij} is the time step of the trip from Z_i to Z_j stated in the agenda, and $v_t(o \to d)$ is the volume of the flow from o to d at time step t. Note that as all the population share the same agenda, the number of people who live in h and work in w is directly given by $v_{t_{01}}(h \to w)$. Note also that as people are required to go home, we have necessarily $v_{t_{01}}(h \to w) = v_{t_{12}}(w \to h)$. Likewise, we have $v_{t_{01}}(h \to w) = v_{t_{23}}(h \to w)$. The probability law given by our model in

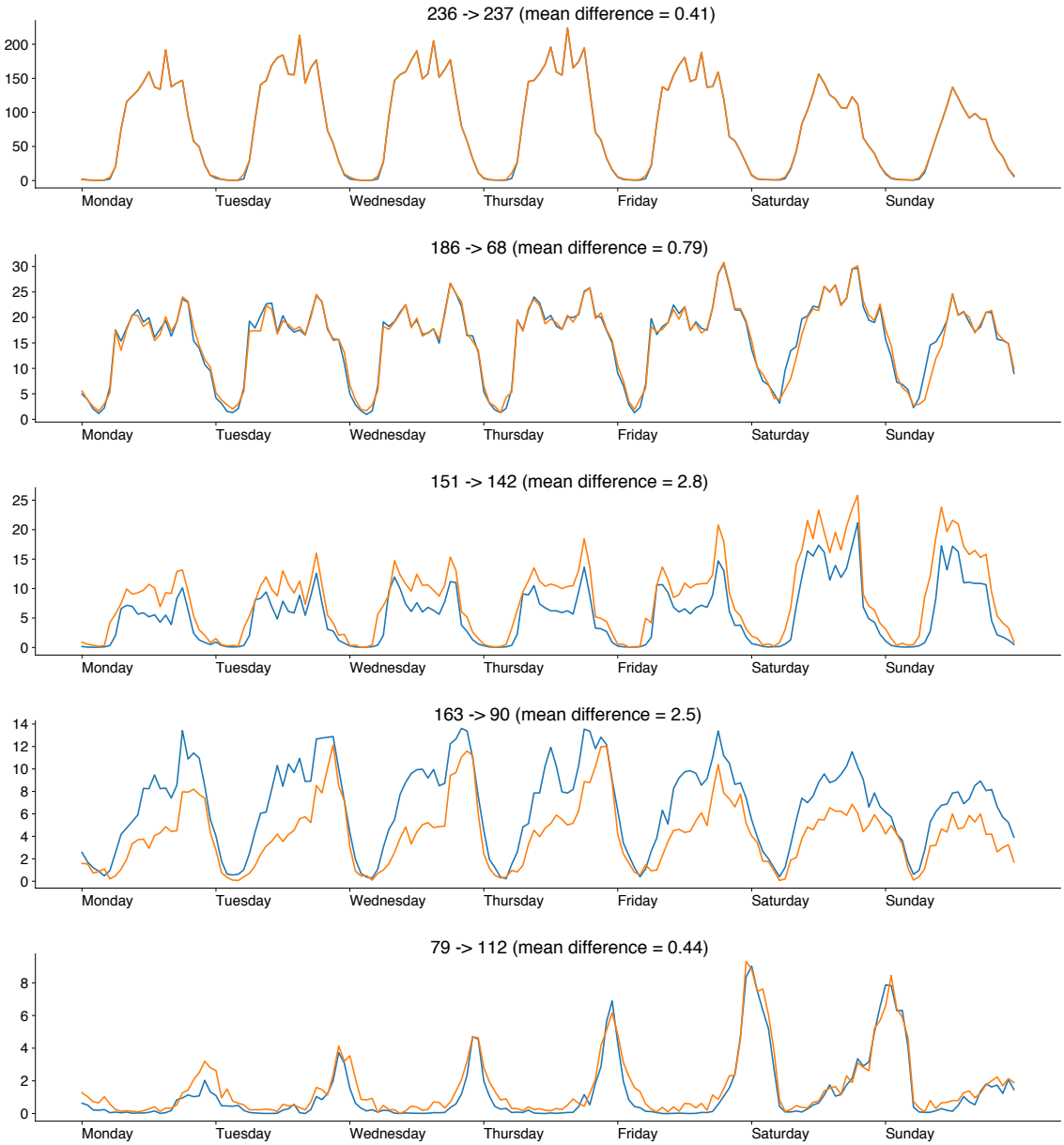


Figure 7.7: Weekly profiles of $o \to d$ flows of various importance in dataset nyc. Blue curve: profile obtained from the OD-matrices anonymised by ATG-dual then reconstructed. Orange curve: profile obtained from the initial OD-matrices.

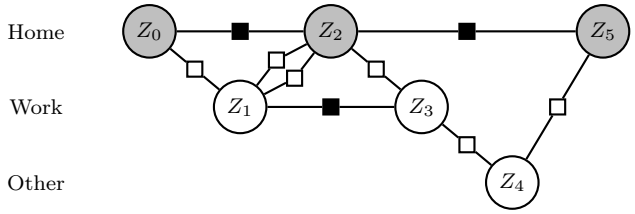


Figure 7.8: Factorisation of the location distribution with all assumptions. White squares denote factors that are conditional probabilities. Black squares are equality factors, forcing their neighboring states to be equal. Note that there are two factors between Z_1 and Z_2 as a result of the factorisation described in Section 5.3.2

Eq. 7.11 then becomes:

$$\hat{p}(Z_1 = Z_3 = w | Z_0 = Z_2 = h, Z_4 = s) \propto v_{t_{01}}(h \to w)^4 \times v_{t_{34}}(w \to s). \tag{7.13}$$

We now express the true proportion of people working in w among those who live in h and go in s for their secondary activity. We will note these proportions as p^* and treat them as probabilities. For ease of reading, in the following we will note the event $Z_0 = Z_2 = h$ simply as h and the event $Z_1 = Z_3 = w$ as w. We are then looking for the proportion $p^*(w|h, Z_4 = s)$. Bayes rule gives us:

$$p^*(w|h, Z_4 = s) = \frac{p^*(Z_4 = s|h, w) \times p^*(w|h)}{p^*(Z_4 = s|h)}, \tag{7.14}$$

where the denominator can be discarded as a constant factor not depending on w. The fraction $p^*(w|h)$ of people working in w among those who live in h is given, from the previous remarks, by:

$$p^*(w|h) \propto v_{t_{01}}(h \to w). \tag{7.15}$$

The other term in Eq. 7.14 is the fraction $p^*(Z_4 = s|h, w)$ of people going to s among those who live in h and work in w. Under the graph model assumption, the independence of the various variables leaves only:

$$p^*(Z_4 = s|h, w) = p^*(Z_4 = s|Z_3 = w) \times p^*(Z_5 = h|Z_4 = s)$$

Only the terms depending on w are of interest here, and $p^*(Z_4 = s|Z_3 = w) \propto v_{t_{34}}(w \to s)$. By simplification of Eq. 7.14, discarding terms that do not depend on h as multiplication constants, we obtain:

$$p^*(w|h, Z_4 = s) \propto v_{t_{01}}(h \to w) \times v_{t_{34}}(w \to s). \tag{7.16}$$

We see that where our model gives a product of several factors depending on h and w (Eq. 7.13), the actual proportion observed depends only on one such term (Eq. 7.16). When all agents have the same agenda, the factors in play here are all equal. If our observations of the v_i are noisy, it makes intuitive sense to retrieve the right dimension of the term by applying a power $\frac{1}{n}$ to the n factors corresponding to transitions between the same locations.

Population with multiple agendas: When the population can have distinct agendas, each used by a sub-population, the remarks made in the previous paragraph are still valid: Indeed,

the graph model assumption states that no matter the total agenda in itself, if all locations are fixed but one, then the probability law for this location is the same. This assumption allows us to keep using the total volume observed as estimators for the factors (Eq. 7.12). Under the same assumption, Eq. 7.15 also holds for the true population. For any given agenda then, the same problem as described in Eq. 7.13 and Eq. 7.16 still holds.

In this discussion, we keep using $v_{t_{ij}}(o \to d)$ as estimator for the factors (Eq. 7.12) and as the true proportions (Eq. 7.15). This is a strong assumption as it seems natural that people would choose their location of shopping on different criteria than their work location. However, it is not a structural assumption: If purpose-dependant OD matrices were available, we could refine the model by using purpose-dependant volumes of the form $v_{t_{ij},p_{ij}}(o \to d)$, without changing anything substantial to the present discussion. The graph model assumption, which is inherent to the probabilistic graph approach and to the use of OD matrices in the spatialisation of trajectories, states that the location of any one activity only depends on the neighborhood of the activity (*i.e.,* the previous one, the next one, and the linked activities) and is independent of the total agenda in itself: this is a rather sensible assumption as the major location choice argument, once an activity has been decided, is indeed the accessibility from the previous activity and to the next one.

7.6 Preprocessing of the agendas

Using transport data from various sources poses a problem of **variability**, as what is described as a trip or an activity may not follow the same definition across sources. For example, a HTS respondent performing two very close activities would declare them as separate in their HTS agenda while they may be detected as one single stay point in the mobile data. Likewise, a short activity would also appear in an HTS agenda but not as a stay point in the mobile data.

More generally, short activities still contribute to the agenda as the trip they are associated with may be longer (and thus more important) than the trips to the following activity. As for activities that are close to each other, they should be merged into a single activity whose purpose is the most important of the group. To achieve this, we perform the following preprocessing for each activity chain. As each activity is associated to a trip in HTS agendas, in the following we associate each trip to its destination activity. Going through each trip featured in a HTS agenda, we either add it to the filtered agenda, discard it, or keep part of the information on the activity as a **candidate** that we will merge with the following trips. The merging logic used to decide among these options is illustrated as a flow chart in Fig. 7.9.

We first check if we have a candidate activity from the processing of the previous trips. If we have no candidate and the destination activity is short (*i.e.,* does not last more than 20 minutes), then we consider that the current trip does not lead to a meaningful activity, although the trip information may be useful. In that case, we store the transport mode and activity start time as a candidate to potentially re-use in the processing of the next trips, and we begin processing the next trip in the input agenda. If the activity lasts more than 20 minutes and the trip distance is more than 100m, then we add the destination activity as the next activity in the filtered agenda. If the trip distance is less than 100m, then we consider that the current activity and the previous one to be one unique activity, whose purpose is

the most important between the two, and the transport mode used is the one used on the longest distance. If on the contrary we do have a candidate from the previous trips, we check once again if the destination activity lasts more than 20 minutes. If not, then we consider the current activity not to be meaningful but we update the trip information of the candidate by keeping the current transport mode if it is used on a longer distance than the candidate's. If the destination activity lasts more than 20 minutes and the trip is longer than 100m, then we add the destination activity as an activity in the agenda after having merged it with the candidate. That is, we keep the start time of the candidate and the mode of the longest trip between the current trip candidate's trip. The activity purpose is the purpose of the current trip. If the current trip is shorter than 100m, then we consider the current activity and the candidate to be one unique activity, whose purpose is the most important purpose between the two. The resulting activity is added to the output agenda.

When merging two activities together, we sometimes have to take the most important purpose of the two. This is done with the following logic: any primary activity is more important than any secondary activity, and in the case of a tie, the longest activity is the most important.

Finally, the filtering of small trips and activities may lead to activity chains featuring trips such as (home $\rightarrow$ home). We simplify the chains by assuming that each agent can only have one home, workplace, and study place, so we merge all consecutive primary activities with the same purpose.

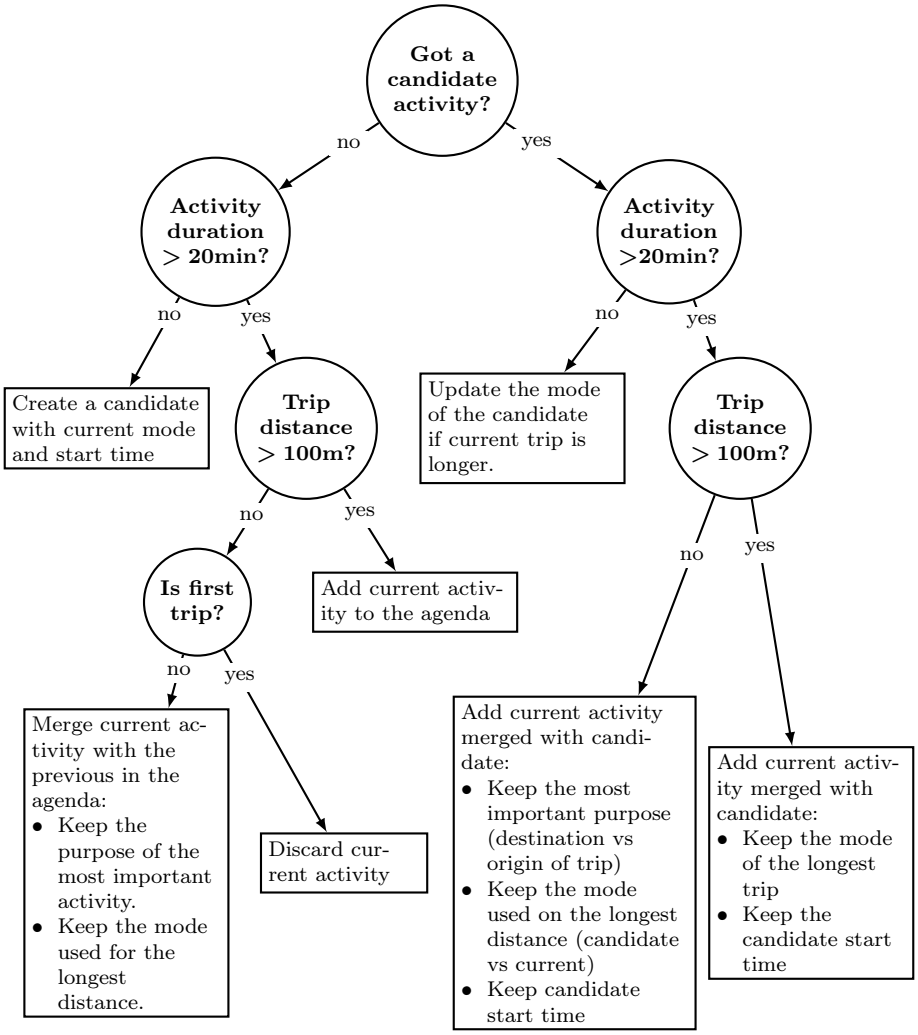


Figure 7.9: Logic of each step of the pre-processing of the HTS agendas. Going through each trip in order, we decide if it must be discarded, added to the agenda as is, merged with the previous trips of the chain, or kept as a **candidate** to be merged with the next trips of the chain.

dataset	solver	time (s)	$\bar{G}$	E	E (normalized)	S
civ	ATG-Dual	00:00:05	57.2	1.72	17.05	10.0%
civ	ATG-Soft	00:00:02	74.4	3.58	18.10	5.3%
civ	oigh	00:00:11	106.8	1.79	17.81	9.9%
civ	glove	00:06:34	35.7	1.01	11.57	10.0%
civ	glove-sk	00:00:09	542.9	1.39	16.81	9.9%
civ	mondrian	00:00:37	52.0	1.32	13.40	0.0%
nyc	ATG-Dual	00:00:01	4.7	0.49	4.75	10.0%
nyc	ATG-Soft	00:00:00	9.1	0.62	6.38	3.3%
nyc	oigh	00:00:05	7.0	0.75	7.05	9.7%
nyc	glove	00:04:42	4.2	0.52	4.61	10.0%
nyc	glove-sk	00:00:09	24.4	0.70	7.48	10.0%
nyc	mondrian	00:01:32	7.5	0.49	6.26	0.0%
senegal	ATG-Dual	00:00:09	17.4	1.21	12.00	10.0%
senegal	ATG-Soft	00:00:04	43.5	1.41	14.28	3.2%
senegal	oigh	00:00:26	26.2	1.36	13.48	9.9%
senegal	glove	04:17:38	16.0	0.82	8.66	10.0%
senegal	glove-sk	00:01:18	667.6	1.21	15.10	10.0%
senegal	mondrian	00:05:33	22.7	0.96	10.6	0.0%
senegal_big	ATG-Dual	00:00:15	3.9	0.17	5.92	9.8%
senegal_big	ATG-Soft	00:00:05	13.9	0.53	8.84	1.1%
senegal_big	oigh	00:00:54	6.2	0.79	10.32	9.5%
senegal_big	glove-sk	00:05:40	80.9	0.68	10.7	10.0%
senegal_big	mondrian	00:32:59	7.5	0.41	7.40	0.0%
senegal_crop	ATG-Dual	00:00:06	13.6	1.19	11.95	10.0%
senegal_crop	ATG-Soft	00:00:01	27.5	1.31	13.53	3.0%
senegal_crop	oigh	00:00:12	32.7	1.35	13.67	9.8%
senegal_crop	glove	01:57:47	13.5	0.82	8.87	10.0%
senegal_crop	glove-sk	00:00:52	508.0	1.04	15.10	10.0%
senegal_crop	mondrian	00:03:43	20.4	0.95	10.63	0.0%
senegal_split	ATG-Dual	00:01:35	16.0	0.75	9.76	10.0%
senegal_split	ATG-Soft	00:00:26	83.5	1.14	13.5	1.8%
senegal_split	oigh	00:15:30	19.7	1.09	12.68	9.9%
senegal_split	glove-sk	00:50:23	2304.5	1.09	14.41	10.0%
senegal_split	mondrian	02:39:21	20.2	0.78	9.66	0.0%

Table 7.1: Detailed results