Université de Montréal

Génération de données: de l'anonymisation à la construction de populations synthétiques

par

Pascal Jutras-Dubé

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Informatique

15 Novembre 2022

Université de Montréal

Faculté des arts et des sciences

Ce mémoire intitulé

Génération de données: de l'anonymisation à la construction de populations synthétiques

présenté par

Pascal Jutras-Dubé

a été évalué par un jury composé des personnes suivantes :

Emma Frejinger
(président-rapporteur)

Fabian Bastin
(directeur de recherche)

Manuel Morales
(codirecteur)

Pierre L'Écuyer
(membre du jury)

Résumé

Les coûts élevés de collecte de données ne rendent souvent possible que l'échantillonnage d'un sous-ensemble de la population d'intérêt. Il arrive également que les données collectées renferment des renseignements personnels et sensibles au sujet des individus qui y figurent de sorte qu'elles sont protégées par des lois ou des pratiques strictes de sécurité et gouvernance de données. Dans les deux cas, l'accès aux données est restreint. Nos travaux considèrent deux angles de recheche sous lesquels on peut se servir de la génération de données fictives pour concevoir des modèles d'analyse où les données véritables sont inaccessibles.

Sous le premier angle, la génératon de données fictives se substitue aux données du recensement. Elle prend la forme d'une synthèse de population constituée d'individus décrits par leurs attributs aux niveaux individuel et du ménage. Nous proposons les copules comme nouvelle approche pour modéliser une population d'intérêt dont seules les distributions marginales sont connues lorsque nous possédons un échantillon d'une autre population qui partage des caractéristiques de dépendances interdimensionnelles similaires. Nous comparons les copules à l'ajustement proportionnel itératif, technologie répandue dans le domaine de la synthèse de population, mais aussi aux approches d'apprentissage automatique modernes comme les réseaux bayésiens, les auto-encodeurs variationnels et les réseaux antagonistes génératifs lorsque la tâche consiste à générer des populations du Maryland dont les données sont issues du recensement américain. Nos expériences montrent que les copules surpassent l'ajustement proportionnel itératif à modéliser les relations interdimensionnelles et que les distributions marginales des données qu'elles génèrent correspondent mieux à celles de la population d'intèrêt que celles des données générées par les méthodes d'apprentissage automatique.

Le second angle considère la génération de données qui préservent la confidentialité. Comme la désensibilisation des données est en relation inverse avec son utilité, nous étudions en quelles mesures le k-anonymat et la modélisation générative fournissent des données utiles relativement aux données sensibles qu'elles remplacent. Nous constatons qu'il est effectivement possible d'employer ces définitions de confidentialité pour publier des données utiles, mais la question de comparer leurs garanties de confidentialité demeure ouverte.

Mots clés : Génération de données, Copules, Synthèse de population, Confidentialité

Abstract

The high costs of data collection can restrict sampling so that only a subset of the data is available. The data collected may also contain personal and sensitive information such that it is protected by laws or strict data security and governance practices. In both cases, access to the data is restricted. Our work considers two research angles under which one can use the generation of synthetic data to design analysis models where the real data is inaccessible.

In the first project, a synthetically generated population made up of individuals described by their attributes at the individual and household levels replaces census data. We propose copulas as a new approach to model a population of interest whose only marginal distributions are known when we have a sample from another population that shares similar interdimensional dependencies. We compare copulas to iterative proportional fitting, a technology developed in the field of population synthesis, but also to modern machine learning approaches such as Bayesian networks, variational autoencoders, and generative adversarial networks when the task is to generate populations of Maryland. Our experiments demonstrated that the copulas outperform iterative proportional fitting in modeling interdimensional relationships and that the marginal distributions of the data they generated match those of the population of interest better than those of the data generated by the machine learning methods.

The second project consists of generating data that preserves privacy. As data privacy is inversely related to its usefulness, we study to what extent k-anonymity and generative modeling provide useful data relative to the sensitive data they replace. We find that it is indeed possible to use these privacy definitions to publish useful data, but the question of comparing their privacy guarantees remains open.

Keywords: Data Generation, Copulas, Population Synthesis, Privacy

Table des matières

Résumé 5

Abstract 7

Liste des tableaux 11

Liste des figures 13

Liste des sigles et des abréviations 15

Remerciements 19

Chapitre 1. Introduction 21

1.1. Construction de populations synthétiques 22

1.2. Confidentialité 23

Chapitre 2. Revue de littérature 25

2.1. Construction de populations synthétiques 25

2.2. Confidentialité 27

Chapitre 3. Copula-based synthetic population generation 29

3.1. Introduction 31

3.2. Literature review 32

3.3. Methodology 35

3.3.1. Bayesian networks 35

3.3.2. Conditional tabular generative adversarial network 36

3.3.3. Tabular variational autoencoder 37

3.3.4. Copula based generation 38

3.4. Numerical experiments 41

3.4.1. Population synthesis for the State of Maryland 43

3.4.2. Population synthesis at the census tract level 46
3.4.3. Learning the dependencies from one PUMA to another 47
3.5. Conclusion 48
Acknowledgments 50
Références bibliographiques 51
Chapitre 4. Définitions de confidentialité 57
4.1. k-anonymat 57
4.2. Données synthétiques 58
4.3. Confidentialité différentielle 60
Chapitre 5. Désensibilisation de Adult Dataset 65
Chapitre 6. Conclusion 71
Références bibliographiques 73

Liste des tableaux

3.1 List of variables 42

3.2 State level SRMSE values between the target population and synthetic agents, column n represents the SRMSE averaged over the possible n-tuples variables. The last column lists sampling zero counts. 44

3.3 County level SRMSE values between the target population and synthetic agents, column n represents the SRMSE averaged over the possible n-tuples variables. The last column lists sampling zero counts. 45

3.4 SRMSE values between the target population and synthetic agents for the PUMA to PUMA experiment, column n represents the SRMSE averaged over the possible n-tuples variables. The last column lists sampling zero counts. 48

4.1 Jeu de données dont les champs "Âge", "Éducation", "Occupation" et "Ville" adhèrent au 2-anonymat. 58

5.1 Description des attributs du jeu de données Adulte. Les variables numériques ont été discrétisées chacune en 10 catégories. 66

Liste des figures

3.1 Schematic description of a variational autoencoder for wich $q_\phi(\cdot \mid x)$ is assumed to be the family of Gaussian distributions. 38

3.2 Empirical cumulative distribution function and its relaxed continuous extension . 41

3.3 Example of a Bayesian network found by the greedy search algorithm. It illustrates a factorization of the joint probability distribution of census variables. 44

3.4 Marginal fit of the county for the state level experiment 45

3.5 Marginal fit of the PUMA for the county level experiment 46

3.6 Marginals fit for tract 702204 when source is PUMA 1201 47

5.1 Score $f1$ pour la prédiction du revenu annuel pour CTGAN et l'algorithme de Mondrian en fonction de différentes valeurs de k 68

5.2 Score d'utilité DWP pour CTGAN et l'algorithme de Mondrian en fonction de différentes valeurs de k 69

Liste des sigles et des abréviations

ACS	Enquête sur la communauté américaine, de l'anglais *American Community Survey*
CDF	Fonction de répartition, de l'anglais *Cumulative Distribution Function*
CTGAN	Réseau antagoniste génératif conditionnel tabulaire, de l'anglais *Conditional Tabular Generative Adverserial Network*
DP	Confidentialité différentielle, de l'anglais *Differential Privacy*
DWP	Prédiction dimensionnelle, de l'anglais *Dimension-wise Prediction*
GAN	Réseau antagoniste génératif, de l'anglais *Generative Adverserial Network*
IPF	Ajustement proportionnel itératif, de l'anglais *Iterative Proportional Fitting*
IPU	Mise à jour proportionnelle itérative, de l'anglais *Iterative Proportional Updating*

IRS	Service du revenu interne, de l'anglais *Internal Revenue Service*
MCMC	Monte Carlo par chaînes de Markov, de l'anglais *Markov Chain Monte Carlo*
MDL	Longueur de description minimale, de l'anglais *Minimum description length*
ML	Apprentissage automatique, de l'anglais *Machine Learning*
PPDP	Publication de données préservant la confidentialité, de l'anglais *Privacy-Preserving Data Publishing*
PUMA	Micro-zone à usage public, de l'anglais *Public Use Micro Area*
QI	Quasi-identifiant
SRMSE	Erreur quadratique moyenne normalisée, de l'anglais *Standardized Root Mean Squared Error*
SZ	Echantillonnage zéro, de l'anglais *Sampling Zero*
TVAE	Auto-encodeur variationnel tabulaire, de l'anglais *Tabular Variational Autoencoder*
VAE	Auto-encodeur variationnel, de l'anglais *Variational Autoencoder*

VGM	Mélange gaussien variationnel, de l'anglais *Variational Gaussian Mixture*

Remerciements

Je remercie particulièrement Fabian Bastin pour sa pédagogie sans égale. Je remercie aussi Cinzia Cirillo pour la collaboration fructueuse et l'invitation à présenter à la conférence SAE 2022 au Maryland. J'aimerais finalement souligner le soutien professionnel de Manuel Morales ainsi que Fin-ML et Mitacs pour leurs supports financiers.

Chapitre 1

Introduction

La modélisation générative peut être décrite comme la tâche qui consiste à apprendre, d'une manière ou d'une autre, des distributions de probabilités jointes sur plusieurs variables dans le but de produire de nouvelles données qui auraient vraisemblablement pu être tirées de l'ensemble de données d'origine.

Le présent mémoire s'intéresse particulièrement à l'application de la modélisation générative en synthèse de population, où ces termes désignent justement la tâche qui consiste à élargir un échantillon représentatif en un recensement entier d'une population d'intérêt. Ce premier projet, en collaboration avec professeur Cinzia Cirillo du département d'ingénierie civile et environnementale de l'université du Maryland, s'est valu la rédaction d'un article scientifique destiné à être publié dans la revue *Transportation Research Part C* que nous présentons dans ce mémoire.

En second lieu, ce mémoire étudie si des données générées synthétiquement peuvent être substituées à des données sensibles pour concevoir des modèles d'analyses de haute qualité, mais qui respectent des contraintes de confidentialité. Cette dimension, moins conséquente relativement à la première, est une extension du travail qui a fait l'objet de recherches dans le cadre d'un stage Mitacs de quatre mois à la Banque nationale du Canada. Les détails de l'implémentation du code, de la méthodologie relative aux processus d'affaire et des données ne peuvent être révélés conformément aux ententes de propriété intellectuelle. Ce mémoire présente néanmoins une revue théorique de méthodes de désensibilisation de données étudiées dans le cadre du projet. Afin de tout de même présenter des résultats, des données publiques ont été désensibilisées pour ce mémoire et on y présente une ébauche d'analyse de qualité des données désensibilisées, ce qui devrait somme toute donner un aperçu des apprentissages acquis durant le stage.

Les sections 1.1 et 1.2 d'introduction motivent les directions de recherche en génération de données dans les contextes respectifs de la synthèse de population et de la publication de données sensibles. Les sections 2.1 et 2.2 dépeignent les états actuels de la littérature

dans ces domaines. La section subséquente présente l'article scientifique en synthèse de population qui détaille particulièrement la méthodologie et les résultats de nos travaux de recherche. Cet article représente la majeure partie du travail académique accompli durant la maîtrise. Des définitions de confidentialité ainsi que leurs garanties théoriques relatives à différents risques s'ensuivent au chapitre 4. Ensuite, le chapitre 5 compare l'utilité de versions désensibilisées de données publiques. Enfin, une conclusion (chapitre 6) résume et discute les deux dimensions de ce mémoire et suggère des directions de recherches ultérieures.

1.1. Construction de populations synthétiques

Les modèles de micro-simulation (Waddell, 2002; Salvini et Miller, 2005) simulent les comportements d'agents au fil du temps afin d'évaluer l'impact de politiques urbaines dans des environnements fictifs, mais réalistes. Ils sont typiquement utilisés en transport pour prédire la demande de voyages entre des régions et prévoir, par exemple, la nécessité de construire de nouvelles routes entre celles-ci, mais sont également utilisés d'autres façons comme pour analyser la demande d'énergie (Panos et Margelou, 2019).

La synthèse de population est le processus par lequel des agents et leurs informations socio-économiques aux niveaux individuel et du ménage sont générés pour différentes régions d'intérêt dans le but d'être donnés comme entrée aux modèles susmentionnés. En fait, les populations synthétiques se substituent aux données du recensement qui sont disponibles en quantité limitée parce que leur collecte est onéreuse et qu'elles sont sensibles du fait qu'elles permettent d'identifier des individus. En général, les gouvernements ne fournissent qu'un sous-ensemble et les distributions marginales de la population totale.

Une approche populaire en synthèse de population consiste à échantillonner des enregistrements de ce sous-ensemble de manière à respecter les distributions marginales de la population cible. Elle correspond bien à la description du problème parce qu'elle utilise à la fois l'échantillon et les informations marginales mis à sa disposition. Toutefois, elle se limite à répliquer des enregistrements connus et ne peut échantillonner d'autres données probables dans la population cible.

En revanche, l'approche probabiliste traite l'échantillon donné comme une réalisation de la population d'intérêt. Modéliser la distribution de probabilité et échantilloner selon cette distribution rend possible la génération de données inconnues mais probables, mais cette approche n'intègre pas la connaissance des distributions marginales de la population cible.

Notre recherche a pour objectif de répondre au questionnement suivant. Est-il possible d'échantillonner des données synthétiques selon l'approche probabiliste tout en encourageant la population générée à respecter les distributions marginales données? Pour y répondre, nous introduisons une méthode – en quelque sorte un "méta-algorithme" – qui tire parti de la théorie des copules pour encadrer les modèles probabilistes afin de faire correspondre

les distributions marginales de la population générée à celles de la population cible. En ce sens, le cadre que nous proposons permet de générer des données pour une population dont seules les distributions marginales sont connues en utilisant un échantillon d'une autre population partageant des dépendances similaires. Nous appliquons ce cadre à des modèles probabilistes apparaissant dans la littérature de population synthétique, soit les réseaux bayésiens, les auto-encodeurs variationnels et les réseaux antagonistes génératifs dont la tâche est de générer des données du recensement américain pour l'état du Maryland. Nous montrons qu'ils bénéficient de notre cadre qui permet d'étudier la structure des données de manière robuste aux particularités des distributions marginales tout en profitant des avantages des modèles sous-jacents.

1.2. Confidentialité

À mesure que les technologies de collecte et de conservation des données électroniques augmentent en puissance, les données recueillies deviennent de plus en plus détaillées et riches, mais aussi de plus en plus sensibles, ce pourquoi elles sont également de plus en plus réglementées. C'est le cas des données du recensement dont l'accès publique est limité à un sous-échantillon de la popoulation parce qu'elles comportent des renseignements sensibles comme le revenu. De même, en santé, les champs personnellement identifiants combinés aux informations médicales suscitent des préoccupations en matière de confidentialité. L'accès à ces données est restreint, ce qui freine les recherches biomédicales et le progrès des soins de patients. Similairement, en vertu de leur faible appétit pour le risque, les institutions financières adoptent des pratiques strictes de sécurité et de gouvernance des données qui en retour ralentissent leurs capacités analytiques internes.

Historiquement, la pratique en matière de publication de données sensibles reposait principalement sur des politiques discriminant les types de données pouvant être publiées et sur des accords contractuels sur l'utilisation et le partage des données. Cette approche limite complètement l'accès à certaines données potentiellement intéressantes et peut inversement fournir des protections insuffisantes. Alternativement, pour atténuer les risques liés à la vie privée et libérer le potentiel des données sensibles, des données homologues désensibilisés peuvent être générées et partagées. C'est ce qu'on désigne par la publication de données préservant la confidentialité (PPDP, de l'anglais *Privacy-Preserving Data Publishing*).

Nos travaux de recherches ont pour but de développer des capacités de publication de données préservant la confidentialité, notamment dans le cas où elles servent à développer des modèles prédicitfs. Les données générées doivent d'une part répondre aux besoins analytiques et s'inscrire dans le processus de livraison de produits d'apprentissage automatique (les besoins d'utilité), d'autre part satisfaire les exigences de sécurité et gouvernance des données (les besoins de confidentialité). Constatons que la protection de la vie privée et la publication

de données exactes et utiles sont des objectifs opposées: le premier voudrait modifier le plus possible les données et le second le moins possible.

En commençant par une identification des approches modernes et d'une étude de leurs garanties respectives de confidentialité, le stage s'est d'abord consacré à l'implémentation d'une solution de désensibilisation de données et ensuite à montrer que les données qu'elle génère conservent une utilité lorsq'on les emploie aux mêmes fins que les données originales. Pour donner suite aux résultats positifs du stage, la question toujours ouverte de l'évaluation du compromis utilité-confidentialité fait aujourd'hui l'objet d'efforts de recherche, ce pourquoi cette dimension n'est pas abordée dans ce mémoire.

Chapitre 2

Revue de littérature

2.1. Construction de populations synthétiques

L'ajustement proportionnel itératif (IPF, de l'anglais *Iterative Proportional Fitting*) (Deming et Stephan, 1940) est largement répandu en synthèse de population en raison de sa simplicité et parce qu'il correspond bien à la description du problème (Duguay *et al.*, 1976; Beckman *et al.*, 1996; Salvini et Miller, 2005; Guo et Bhat, 2007; Auld et Wies, 2009; Ye *et al.*, 2009). Initialement décrit comme méthode numérique générale pour ajuster une matrice qui soit la plus proche possible d'une matrice initiale, mais qui respecte les totaux de lignes et colonnes d'une matrice cible, IPF répond au problème d'ajustement qui se pose dans le recensement d'une population où un échantillon et un décompte complet de ses charactéristiques marginales sont disponibles. Pour générer une population synthétique avec ces données, IPF nécessite deux étapes. Dans l'étape d'ajustement, un tableau de contingence est calculé à partir de la population source et des distributions marginales. Dans l'étape d'allocation, des enregistrements de la population source sont échantillonnés aléatoirement selon les poids donnés par le tableau de contingence.

Choupani et Mamdoohi (2016) décrivent les limitations de IPF et révèlent que le problème de cellule nulle (Guo et Bhat, 2007) est particulièrement critique. Lorsqu'un groupe démographique de la population cible n'est pas représenté dans l'échantillon source, ses poids demeurent nuls dans le tableau de contingence de sorte qu'il ne peut être échantillonné à l'étape d'allocation. Pour y remédier, Guo et Bhat (2007) proposent d'attribuer des poids initiaux aux groupes démographiques absents de l'échantillon source, mais cette solution requiert une connaissance préalable du domaine. En d'autres mots, la population générée par IPF n'est que constituée de réplications provenant de l'échantillon source et les combinaisons qui y sont absentes manquent d'être représentées dans la population cible.

Une autre veine de travaux s'intéresse aux méthodes de simulation Monte Carlo par chaînes de Markov (MCMC, de l'anglais *Markov Chain Monte Carlo*). Le principe de l'approche MCMC est d'utiliser l'échantillonneur de Gibbs qui visite séquentiellement les variables et échantillonne leurs valeurs en exploitant leurs distributions conditionnelles. Le défi est de spécifier ces distributions conditionnelles. Farooq *et al.* (2013) proposent de les estimer par les fonctions de répartition conditionnelle empiriques, mais cette approche souffre également du problème de cellule nulle. De plus, l'échantillonnage de Gibbs peut ne jamais visiter certaines combinaisons lorsqu'il n'existe pas de chemins entre des îles d'états et ce problème s'exprime plus conséquemment lorsque le nombre de dimensions augmente. Les réseaux bayésiens (Sun et Erath, 2015) évitent intelligemment ces difficultés en abstrayant la structure des données à l'aide d'un graphe orienté acyclique et des probabilités conditionnelles locales. Ils optimisent la structure du graphe pour obtenir des représentations appropriées et interprétables des distributions conditionnelles.

Alternativement, le problème de synthèse de population peut être décrit comme une tâche d'apprentissage automatique qui consiste à apprendre des distributions de probabilités jointes multivariées. Garrido *et al.* (2020) utilisent les autoencodeurs variationnels (VAEs, de l'anglais *Variational Autoencoder*) (Kingma et Welling, 2014) et les réseaux antagonistes génératifs (GANs, de l'anglais *Generative Adverserial Network*) (Goodfellow *et al.*, 2014) pour générer des populations synthétiques. Les VAEs sont des modèles génératifs dans lesquels l'entrée est encodée en une distribution latente multivariée puis décodée par le modèle génératif qui apprend à reconstruire l'entrée. Les GANs sont une autre classe d'algorithmes génératifs d'apprentissage automatique dans lesquels un générateur s'oppose à un discriminateur qui apprend à déterminer si un échantillon provient de la distribution du modèle ou de la distribution des données. Ces travaux montrent que les modèles génératifs profonds sont adaptés au problème de synthèse de population en grandes dimensions. Les modèles génératifs d'apprentissage automatique adressent naturellement le problème de cellule nulle parce qu'ils évaluent, explicitement ou non, la distribution de probabilités jointes et permettent le prélèvement d'échantillons à partir de la distribution apprise.

Une autre perspective qui gagne rapidement en popularité dans les domaines nécessitant la modélisation de données multivariées est celle des copules. On trouve notamment les travaux de Bhat et Eluru (2009) qui étudient les effets de l'environnement sur les habitudes de déplacements et distances parcourues quotidiennement par les ménages; de Born *et al.* (2014) qui modélisent la participation à des évènements discrétionnaires pour prédire les demandes de déplacements la fin de semaine; et de Jeong *et al.* (2016) qui comparent IPF aux copules quant à leurs capacités de respecter les distributions jointes des données de référence.

Formalisée par Sklar (1959), la copule est une fonction de répartition multivariée pour laquelle la distribution de probabilité marginale de chaque variable suit une loi uniforme

standard. Elles ont comme propriété remarquable de caractériser les dépendenses interdimensionnelles sans subir les effets des lois marginales. En effet, le théorème de Sklar (1959) stipule que toute distribution conjointe multivariée peut être écrite en termes de fonctions de distributions marginales univariées et d'une copule qui décrit la structure de dépendance entre les variables. Il montre également l'unicité des copules continues, mais l'unicité ne tient pas nécessairement dans le cas des copules discrètes. D'ailleurs, le théorème de Sklar ne décrit en rien comment trouver une copule appropriée. Des familles paramétriques de copules comme les copules indépendante, archimédienne et gaussienne peuvent être postulées au risque de limiter la justesse de la modélisation des dépendances.

Nous proposons d'apprendre une copule non-paramétrique et d'interpréter les données générées par la copule apprise avec les informations marginales de la population cible pour prendre à la fois avantage de la complexité des modèles probabilistes décrits dans la littérature de population synthétique et de la capacité des copules à séparer les dépendances interdimensionnelles des distributions marginales.

2.2. Confidentialité

Le concept de publication de données préservant la confidentialité (PPDP) induit naturellement à celui de la dépersonnalisation. Un renseignement concernant une personne physique est dépersonnalisé lorsqu'il ne permet plus d'identifier directement la personne concernée. Pour qu'un jeu de données soit dépersonnalisé, il suffit que les champs directement identifiants comme les prénoms, noms de famille, numéros d'assurance sociale, etc. y soient masqués ou supprimés. La dépersonnalisation ne se préoccupe pas des champs dits quasi-identifiants qui à eux seuls ne peuvent pas identifier la personne concernée, mais qui le peuvent lorsque combinés. de Montjoye *et al.* (2015) ont montré que 80% du total des consommateurs pouvaient être réidentifiés par seulement 3 transactions par carte de crédit desquelles uniquement le commerçant et la date de transaction étaient révélés. Il est également possible de faire correspondre des champs quasi-identifiants à ceux de données auxiliaires pour faire ce qu'on appelle une attaque par liaison. Sweeney (2002) découvre le dossier médical du gouverneur du Massachussetts en pairant des données publiques du bureau de vote aux données médicales dépersonnalisées publiées par le gouvernement. Narayanan et Shmatikov (2008) ont retrouvé des utilisateurs de Netflix en comparant les classements et les horodatages des données publiques de l'*Internet Movie Database* et des données dépersonnalisées publiées dans le cadre d'un défi sur les systèmes de recommandations lancé par Netflix.

Un détenteur de données qui veut divulguer une version désensibilisée de ses données est plus prudent s'il s'assure que les méthodes de désensibilisation qu'il emploie sont soutenues par des garanties mathématiques de confidentialité. Le k-anonymat (Sweeney, 2002)

est la propriété d'un jeu de données selon laquelle chaque enregistrement partage les quasi-identifiants d'au moins $k-1$ autres enregistrements. Cette propriété peut être atteinte en partitionnant les données puis en généralisant les champs quasi-identifiants au sein des classes d'équivalence, mais le problème de trouver le partitionnement multidimensionnel qui conserve le plus d'utilité est un problème NP-difficile (Meyerson et Williams, 2004). De nombreuses heuristiques de partitionnement ont été proposées dans la littérature du k-anonymat (Iyengar, 2002; Lin et Wei, 2008; El Emam *et al.*, 2009), la plus connue étant probablement l'algorithme Mondrian (LeFevre *et al.*, 2006), algorithme glouton qui partitionne les données récursivement selon les champs ayant les plus grands diamètres normalisés.

Une approche alternative pour la PPDP est de remplacer les données sensibles par des données synthétiques. Par exemple, medGAN (Choi *et al.*, 2017) adapte un GAN aux données tabulaires en santé en apprenant au préalable une représentation continue des données avec un auto-encodeur. Cependant, les modèles génératifs à eux seuls ne procurent aucune garantie de confidentialité a priori. Dans un contexte de sur-apprentissage, il est possible que les données synthétiques révèlent de l'information sur les données d'entraînement (Bellovin *et al.*, 2018). La confidentalité différentielle (DP, de l'anglais *Differential Privacy*) (Dwork *et al.*, 2006; Dwork, 2011) est une définition mathématique qui garantit au moyen de randomisation que toute séquence de résultats (réponses aux requêtes faites à une base de données) est essentiellement également susceptible de se produire indépendamment de la présence ou l'absence de tout individu. Il est possible d'entraîner un réseau de neurones de manière à ce que ses poids satisfassent la DP en exploitant les propriétés de composition de cette dernière (Abadi *et al.*, 2016; Mironov, 2017). Par conséquent, il est possible d'entraîner des GANs qui respectent la DP (Xu *et al.*, 2019a) et ainsi mesurer la confidentialité de données synthétiques.

Chapitre 3

Copula-based synthetic population generation

par

Pascal Jutras-Dubé[2], Mohammad Bilal Mohammad Al-Khasawneh[1],
Zhichao Yang[1], Javier Bas[3], Fabian Bastin[2], Cinzia Cirillo[1]

([1]) Department of Civil and Environmental Engineering, University of Maryland, College Park, MD, USA

([2]) Department of Computer Sience and Operations Research, Université de Montréal, Montreal, QC, Canada

([3]) Department of Economics, Universidad de Alcalá, Facultad de Ciencias Económicas, Madrid, Spain

Cet article est en préparation et est destiné à être soumis à Transportation Research Part C.

Les principales contributions de Pascal Jutras-Dubé à cet article sont

- L'implémentation de l'entièreté du code;
- Des discussions importantes sur la méthodologie et les résultats, en particulier sur le problème des copules discrètes, les métriques d'évaluation et le choix des expériences numériques;
- Les expériences;
- Les descriptions du réseau bayésien, de l'autoencodeur variationnel et du réseau antagoniste génératif;
- la rédaction des résultats en collaboration avec Javier Bas et Cinzia Cirillo

L'idée fondamentale revient à Fabian Bastin. Javier Bas a majoritairement travaillé sur l'introduction avec Cinzia Cirillo, sur la revue de littérature ainsi que sur la cohérence du style entre les parties. Mohammad Bilal a préparé les données et Zhichao Yang a fait une ébauche de la conclusion.

Résumé. La synthèse de population consiste à générer des représentations synthétiques mais réalistes d'une population cible de micro-agents à des fins de modélisation et de simulation. Nous introduisons une nouvelle approche basée sur les copules pour générer des données synthétiques pour une population cible dont seules les distributions marginales empiriques sont connues en utilisant un échantillon d'une autre population partageant des dépendances marginales similaires. Cette approche permet d'inclure une composante spatiale dans la génération de la population synthétique et de combiner diverses sources d'information pour obtenir des générateurs plus réalistes. Plus précisément, nous normalisons les données et les traitons comme des réalisations d'une copule donnée, entraînons un modèle génératif sur les données normalisées puis injectons les informations marginales. Nous comparons les copules à IPF et aux approches probabilistes modernes telles que les réseaux bayésiens, les auto-encodeurs variationnels et les réseaux antagonistes génératifs et illustrons sur les données du recensement américain que la méthode proposée permet d'étudier la structure des données à différents niveaux géographiques d'un façon robuste aux particularités des distributions marginales.
Mots clés : Synthèse de population, Copule, Apprentissage automatique

Abstract. Population synthesis consists of generating synthetic but realistic representations of a target population of micro-agents for the purpose of behavioral modeling and simulation. We introduce a new framework based on copulas to generate synthetic data for a target population of which only the empirical marginal distributions are known by using a sample from another population sharing similar marginal dependencies, making it possible to include a spatial component in the generation of population synthesis and to combine various sources of information to obtain more realistic population generators. Specifically, we normalize the data and treat them as realizations of a given copula, and train a generative model on the normalized data, before injecting the information on the marginals. We compare the copulas framework to IPF and to modern probabilistic approaches such as Bayesian networks, variational auto-encoders and generative adversarial networks. We also illustrate on American Community Survey data that the method proposed makes it possible to study the structure of the data at different geographical levels in a way that is robust to the peculiarities of the marginal distributions.
Keywords: Population synthesis, Copula, Machine Learning

3.1. Introduction

Population synthesis refers to models that aim at constructing artificial datasets whose characteristics mimic that of a population of interest. Population synthesizers can produce a set of agents with detailed sociodemographic and socioeconomic information at both the individual and household levels, maintaining the structural coherence of the population they replicate. Therefore, these methodologies have been of great use wherever an agent-based model was to be implemented, such as transportation models based on micro-simulation Arentze et al. (2007); Pritchard and Miller (2012); Müller and Axhausen (2011); Guo and Bhat (2007), models in which analyzing the spatial implications of a given policy is capital.

In order to properly reflect the population properties it aims to reflect, a synthetic population must share the same joint distribution of the variables that it encapsulates. In the context of microsimulation frameworks for travel behavior, this property must hold for the geographic area of interest. Various techniques have been proposed to achieve this goal. However, with a few notable exceptions, as Barthelemy and Toint (2013), most of them rely on a sample of the target population (e.g. census data, travel survey), often costly to obtain. As a result, the sample size can be limited, especially when working at small geographical level (e.g. census tract level). It is therefore desirable to include a spatial component in the population synthesis generation and to combine various information sources to obtain more realistic population generators. However, such aspects of the problem have received very limited attention in the transportation literature. Thus, we here introduce a framework that can generate synthetic data for a target population of which only the marginal or empirical marginal distributions are known at different geographical levels (e.g. state, county, Public Use Micro Areas [PUMA], census tract) by using a sample from another population sharing similar marginal dependencies. To achieve this objective, we combine the theory of copulas with traditional machine learning (ML) generative modeling approaches, to separate the learning of the dependencies structure from the marginal distributions, allowing its use with various populations that differ on the marginals only. We compare several methods (Bayesian networks, variational autoencoders, generative adversarial networks) alone or in combination with copulas, and illustrate on American Community Survey (ACS) data that the method proposed makes it possible to study the structure of the data at different geographical levels in a way that is robust to the peculiarities of the marginal distributions.

The rest of the paper is organized as follows. In Section 2, we review the literature on population synthesis, with a focus on the different existing methodologies recently proposed. Section 3 describes the methods applied in this work, while the specifics of the experiments carried out, as well as the results obtained, are detailed in Section 4. Finally, conclusions and future research avenues are presented in Section 5.

3.2. Literature review

The ultimate goal of a population synthesizer is to use in different processes (modeling, optimization, simulation, etc.) the new information that it generates from an available disaggregated dataset insufficient in size. There are several notable examples in transportation where this exercise is performed. Eluru et al. (2008) develop a microsimulator called CEMUS that feeds aggregated socioeconomic information into a population generator to produce household and individual disaggregated synthetic datasets to model individual-level activity-travel patterns. Bradley et al. (2010) present a forecasting model implemented by the Sacramento Area Council of Governments that relies upon the simulation of residents' full-day activity and travel schedule. Auld et al. (2012) use a dynamic simulator to feed the Transportation Analysis and Simulation System (TRANSIMS). Smith et al. (1995) in order to generate route choices. Ziemke et al. (2018) use a synthetic population to compute accessibility measures in Nelson Mandela Bay in South Africa by using census and travel survey data.

A desirable property of any population synthesizer is to preserve the general characteristics of the population, which leads to different technical approaches. The most widely used, until recently, has probably been the Iterative Proportional Fitting (IPF), initially proposed by Deming and Stephan (1940), and popularized in transportation by Duguay et al. (1976). IPF selects households from the source sample trying to match some given marginals totals, requiring a fitting stage and an allocation stage. In the fitting step, a contingency table is computed from the seed table (the source sample) and the marginals totals. In the allocation phase, households are randomly selected from the seed table to match the frequency given in the contingency table. There is an abundant literature on empirical applications of IPF, reviewed for instance by Müller and Axhausen (2011), as well as on its technical aspects (Arentze and Timmermans, 2004; Salvini and Miller, 2005; Auld et al., 2009; Ye et al., 2009; Guo and Bhat, 2007). The method nevertheless presents several important flaws. In this regard, Pritchard and Miller (2012) address computational memory restrictions, while Guo and Bhat (2007) explore how to avoid sampling zero issues. Ye et al. (2009) consider simultaneously fitting different types of agents, proposing a heuristic approach called Iterative Proportional Updating (IPU) to overcome the disadvantages of the standard IPF. However, they fail to accommodate the new synthetic information at multiple geographical resolutions simultaneously, leading to a loss of representativeness. Konduri et al. (2016) extend their efforts, proposing an enhanced IPU algorithm that accounts for constraints at different levels of spatial resolution when generating a synthetic population.

However, as Farooq et al. (2013) point out, fitting a contingency table to the available data may entail errors if the information is not complete or has been manipulated. In fact, these potential errors cannot be contrasted due to absence of a complete real data set. Thus,

these authors propose a Markov Chain Monte Carlo (MCMC) simulation-based approach that uses partial views of the joint distribution of the real population obtained from the census to draw from it high-dimensional synthetic populations that would otherwise have been impossible to produce using IPF. Other authors relying on MCMC are Casati et al. (2015), Saadi et al. (2016b), and Saadi et al. (2016a).

Another category of methods for synthesizing populations, the Combinatorial Optimization Methods, proceed differently. They divide an area into mutually exclusive subareas where a distribution of the attributes of interest is available. Then a sample taken over the whole population is fitted to the given set of marginals for each subarea. Huang and Williamson (2001) offer a comparison of these two approaches in the context of the creation of small-area microdata. However, the assumptions about the data that these two families of methods require cannot always be met. This led Barthelemy and Toint (2013) to develop a new type of generator, applied to the case of Belgium. Their technique works in a three-step process. First, a pool of individuals pertaining to a certain household is generated; then, the household joint distribution is estimated and stored in a contingency table; finally, the synthetic households are constructed by randomly drawing individuals from the pool of individuals, preserving the distribution computed in the second step. They compare the results of this method with those of the extended IPF described in Guo and Bhat (2007), observing that the new generator performs better. Another effort to offer an alternative approach is that of Sun et al. (2018), who propose a hierarchical mixture model to generate representative household structures in population synthesis. Their framework comprises a probabilistic tensor factorization, a multilevel latent class model, and a rejection sampling process. They test their procedure on the Household Interview Travel Survey data of Singapore, being able to generalize the associations among attributes as well as to reproduce structural relationships among household members.

Yet, it is possible to observe a recent increasing trend in the substitution of iterative fitting algorithms for the generation of synthetic populations by alternative approaches. One of these new perspectives is the application of Copula Generative Models. Copulas are a mathematical probability tool that allows the modeling of random variables that have an intrinsic relationship to each other. The dependence structure can be decoupled from the original available information so any set of new synthetic data can be created having that same structure. We refer to Genest and Favre (2007) for a gentle introduction, and to the books by Joe (1997, 2015) and Nelsen (2006) for readers seeking concrete mathematical proofs, theorems and derivations related to copulas. The selection of a specific copula can remain challenging Kaushik et al. (2019), especially in the case of discrete copula. Avramidis et al. (2009) provide efficient correlation matching for modeling dependence of discrete multivariate distribution via normal copula. Copulas have been successfully applied in various domains, including finance and actuarial sciences Cherubini et al. (2004), transportation Bhat and

Eluru (2009); Pinjari et al. (2009); Rana et al. (2010); Kao et al. (2012); Born et al. (2014), water resources management Borgomeo et al. (2015), and simulation modeling of arrival rates in call centers Oreshkin et al. (2016); Jaoua et al. (2013).

Bayesian network (BN) is another alternative approach that has recently gained momentum as a tool to perform population synthesis. This method encodes the dependency relationships among predictors using a graphical model in which nodes represent variables, links represent their conditional dependencies, and probability distributions are assigned to each node, conditional on its parents. As expressed in the seminal work of Sun and Erath (2015), if the conditional structure of the data generating process is known, inference can be based on the underlying probabilities and the population can then be generated accordingly by sampling from the joint distribution. Since this information is typically not available, the authors propose that the graph structure of the data may be learned through a scoring approach. More recent examples of the use of BN for population synthesis are given by Zhang et al. (2019), who make use of traditional survey data and digital records of networking and human behavior to generate connected synthetic populations using BN; Hörl and Balac (2021), from their part, generate a synthetic travel demand based on open data.

On the other hand, other approaches to the generation of synthetic populations fall more along the lines of ML. The most popular of these methodologies are probably generative models (Bishop, 2006), given their success in generating artificial images. A generative model captures the dependency structure of the variables of a dataset and trains a generator to reproduce samples preserving their joint distribution. Among generative models, two families stand out: variational autoencoders (VAE) (Kingma and Welling, 2014) and generative adversarial networks (GAN) (Goodfellow et al., 2014). VAE are unsupervised generative models composed of an encoder and a decoder that are trained to minimize the reconstruction error between the encoded/decoded data and the initial data. When the model is trained, it learns the joint distribution of the data. A conditional VAE was used to study travel preference dynamics using a synthetic pseudo-panel approach in Borysov and Rich (2021), while Boquet et al. (2020) propose a VAE model to generate traffic data. In particular, we use in this paper Tabular VAE (TVAE) (Xu et al., 2019), an adaptation of regular VAE to tabular data. GANs were first introduced by Goodfellow et al. (2014) as a framework in which two neural networks, called generator (which produces the target output) and discriminator (which distinguish true data from the output of the generator), compete with one another to deceive and not be deceived. This process leads to a model that captures the distribution and dependency of the features of a dataset, and ultimately creates new synthetic information. Despite its great success in the computer vision and natural language processing fields, there are only a few attempts of using GAN in transportation. Yazdizadeh et al. (2021) used GAN to infer travel mode from GPS trajectory data and derive trip information from travel survey. The work of Günthermann et al. (2020) delves along the same lines. On the contrary,

Yin et al. (2018) focus on the use of deep generative models to generate synthetic travelers' mobility patterns that replicate the statistical properties of a sample of actual travelers. Given the structure of our data, we use in this paper, as in the case of VAE, a particular GAN. Namely, the so-called Conditional tabular GAN (CTGAN) (Xu et al., 2019). It was designed to adapt GAN to tabular data with a mix of potentially highly imbalanced discrete columns and multimodal non-Gaussian-like continuous columns.

3.3. Methodology

Consider a sample of size N from a random vector X, from which we want to generate a synthetic population of any arbitrary size. As pointed in the literature review, many techniques can be used to generate a synthetic population, but we will select a few of them only, in addition to IPF, capitalizing on the recent results to identify the most promising approaches.

We first briefly introduce Bayesian networks, conditional tabular generative adversarial network (CTGAN), and tabular variational autoencoders (TVAE) (Xu et al., 2019), that will serve as our benchmark methods. We then discuss copulas and their link to normalization procedures in ML, and conclude with their use for synthetic populations generation.

3.3.1. Bayesian networks

A Bayesian network (BN) is a probabilistic model that represents the random variables X_i, $i = 1, \ldots, d$ and their conditional dependencies in the form of a directed acyclic graph G. Each node corresponds to one random variable, and the arcs express the conditional dependencies between them. If no path exists between two nodes, the corresponding random variables are independent. The joint probability distribution of X can be decomposed as the product of the marginal distributions of X_i, conditioned on its parents Π_i, $i = 1, \ldots, d$ using the chain rule

$$P(X \mid G, \Theta) = \prod_{i=1}^{d} P(X_i \mid \Pi_i, \theta_i),$$

where θ_i are the parameters of the distribution of X_i and $\Theta = \{\theta_1, \ldots, \theta_d\}$.

It is possible to learn both the graph structure G and the parameters Θ. This process is often referred to as structural learning and has two stages: model selection and model optimization. In the selection stage, a score function is used to quantify how well a hypothetical graph structure fits the data. For example, the minimum description length (MDL) is a popular scoring function consisting of two components that estimate the structural complexity and the likelihood of the data given the model (Lam and Bacchus, 1994). The goal of the model optimization stage is to identify the hypothetical structure with the highest score.

However, the evaluation of all possible graph structures has unrealistic time complexity (Cooper, 1990). Instead, we can apply a heuristic search to find a convenient structure that greedily chooses a topological ordering of the variables, and optimally identifies the best parents for each variable given this ordering (Heckerman et al., 1995), as illustrated for instance in Figure 3.3, depicting a Bayesian network built on our dataset.

Once the structural learning is completed, a synthetic population can be drawn from the the factorized joint probability distribution defined by the Bayesian network. For more details, we refer the reader to Sun and Erath (2015).

3.3.2. Conditional tabular generative adversarial network

Generative adversarial networks (GANs) (Goodfellow et al., 2014) are a class of unsupervised ML algorithms in which a generative model is opposed to a discriminative model that learns to determine whether a sample comes from the model distribution or from the data distribution. The competition causes both parties to improve until the generative model can generate false data that is indistinguishable from the real data.

The generator G_θ is a neural network with parameters θ that approximates the distribution of data p_X by generating synthetic samples $G_\theta(Z)$ from a prior random distribution $Z \sim p_Z$. The discriminator D_ϕ is a neural network with parameters ϕ whose scalar output $D_\phi(x)$ represents the probability that an example x comes from the real data rather than from the distribution of the generator. D_ϕ is trained to maximize the probability of assigning the correct class whether an example was sampled from the training set or from G_θ. G_θ is simultaneously trained to minimize the probability of D_ϕ assigning the correct class to its output

$$\min_\theta \max_\phi \mathbb{E}_{X \sim p_X}[\log D_\phi(X)] + \mathbb{E}_{Z \sim p_Z}[\log(1 - D_\phi(G_\theta(Z)))].$$

GANs have achieved impressive performance in generating high-quality synthetic images but were not explicitly designed to manipulate tabular data. This led to the development of conditional tabular GANs (CTGAN) (Xu et al., 2019) that can handle on tabular data with a mix of potentially highly imbalanced discrete columns and multimodal non-Gaussian-like continuous columns.

A CTGAN uses mode-specific normalization to deal with columns with complicated distributions, which is often the case for continuous features in tabular data. For a continuous column, mode-specific normalization first estimates the number of modes with a variational Gaussian mixture model (VGM). Then, for each value in the column, it calculates the probability that it comes from each mode with the distributions given by the VGM. Finally, it replaces the value with a mode picked according to the computed probabilities and a scalar representing the value within the mode.

A CTGAN also integrates a conditional generator into the architecture of a GAN to manage the class imbalances of categorical columns, by leveraging three key elements: the conditional vector, the generator loss, and the training-by-sampling method. The conditional vector indicates if a particular discrete column D_{i^*} must have a particular value k^*. Since the D_{i^*} is represented as a one-hot encoding d_{i^*}, the condition can be written $d_{i^*}^{(k^*)} = 1$. Each discrete column D_i has a corresponding mask vector m_i shaped like its one-hot encoding d_i to represent the condition

$$m_i^{(k)} = \begin{cases} 1 & \text{if } i = i^* \text{ and } k = k^* \\ 0 & \text{otherwise.} \end{cases}$$

The conditional vector is the concatenation of these masks. During the feed-forward pass, nothing encourages the conditional generator to produce the desired output for D_{i^*} for which there is the condition $D_{i^*} = k^*$. The mechanism proposed to enforce the conditional generator to respect the condition is to penalize its loss by adding the cross-entropy between the generated discrete columns and the conditional vector. The training-by-sampling approach samples the conditional vector and training data to help the model evenly explore all possible values in discrete columns. It first randomly selects a discrete column, then constructs a probability mass function across the range of values for that column and samples a value according to this mass function. Finally, it creates the conditional vector representing the sampled column and value. We refer to Xu et al. (2019) for a deeper dive into CTGAN's formalism.

3.3.3. Tabular variational autoencoder

The architecture of a variational autoencoder (VAE) (Kingma and Welling, 2014), considered by Borysov et al. (2019) in the context of synthetic populations, is analogous to an autoencoder. A probabilistic encoder $q_\phi(Z \mid X)$ is meant to map the input to a multivariate latent distribution and a probabilistic decoder $p_\theta(X \mid Z)$ maps back the latent distribution to the data space. The probabilistic encoder and decoder are represented by neural networks, with parameters ϕ and θ, respectively.

We would like to approximate the data distribution by first sampling from a known prior distribution. Let z represent a latent encoding of a data point x and consider their joint distribution $p_\theta(X,Z)$. Conditioning on Z gives us the parameterized approximation $p_\theta(X)$ of the data distribution:

$$p_\theta(x) = \int_z p_\theta(x \mid z) p_\theta(z) dz.$$

The computation of $p_\theta(x)$ is usually intractable, for instance like when likelihood functions $p_\theta(x \mid z)$ correspond to neural networks with nonlinear hidden layers. For this reason, we need to introduce an auxiliary parameterized family of functions, typically the Gaussian

distributions $\mathcal{N}(\mu(x),\Sigma(x))$, to approximate the posterior distribution

$$q_\phi(z \mid x) \approx p_\theta(z \mid x).$$

as schematized in Figure 3.1. We want to make $q_\phi(z \mid x)$ as close as possible to $p_\theta(z \mid x)$ by minimizing a distance measure between distributions. In addition, we would like to optimize the generative model to reduce the reconstruction error between the input and output. Maximizing the evidence lower bound (ELBO) defined as

$$\begin{aligned} L(x) &= \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\frac{p_\theta(x,z)}{q(z \mid x)}\right] \\ &= \ln p_\theta(x) - D_{KL}\left(q(z \mid x) \parallel p_\theta(z \mid x)\right). \end{aligned}$$

is equivalent to minimizing the reverse Kullback-Leibler divergence between the approximation and true posteriors and jointly maximizing the log-likelihood of the data, and is typically done using a stochastic gradient descent procedure. For more details, see Kingma and Welling (2014).

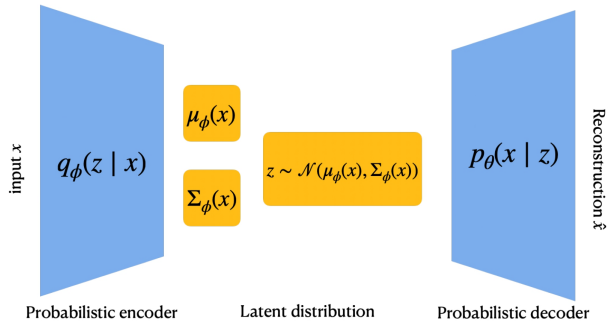


Fig. 3.1. Schematic description of a variational autoencoder for wich $q_\phi(\cdot \mid x)$ is assumed to be the family of Gaussian distributions.

A tabular variational autoencoder (TVAE) (Xu et al., 2019) adapts a VAE to tabular data with the same preprocessing and loss function modification procedures as in the CTGAN framework.

3.3.4. Copula based generation

Most of household surveys and similar studies rely on detailed, but small samples, while we can access to more complete information regarding specific aspects. In particular, data from household surveys are typically presented as multivariate vectors implicitly capturing the dependencies between their components, but we have sometimes access to the distributions of individual marginals at the population level. Another typical situation is when we want to focus on a sub-area covered by the sample, resulting in a limited number of observation vectors, while we have a comprehensive knowledge of the distribution of the individual factors. Finally, we are sometimes in areas where we have access to the marginal

distributions only, and no household vector data. For example, in a census data generation context, it is reasonable to conjecture that marginals encode demographic elements and that multidimensional dependencies encode more complex socioeconomic patterns shared across related regions. Assuming that the dependencies can be adequately captured by a sample over a different geographical sector, we aim to be able to exploit this dependencies structure over the area of interest. We thus need a tool to separate the dependencies structure to the specific marginal distributions, which is provided by the copula theory.

A d-dimensional copula C is a multivariate cumulative distribution function (CDF) on $[0,1]^d$ having all marginals uniformly distributed on $[0,1]$ (Nelsen, 2006; Joe, 1997, 2015; Okhrin et al., 2017). A fundamental result, due to Sklar (1959), states that any multivariate distribution can be represented by means of a copula.

Théorème 1 (Sklar's theorem)**.** *Let H be a multivariate distribution function with margins $F_1, \ldots, F_d$, then there exists a copula C such that*

$$H(x_1, \ldots, x_d) = C\left(F_1(x_1), \ldots, F_d(x_d)\right), \quad x_1, \ldots, x_d \in \overline{\mathbb{R}}. \tag{3.3.1}$$

If F_j, $j = 1, \ldots, d$, are continuous then C is unique. Otherwise C is uniquely determined on the Cartesian product of the range of the marginals $F_1\left(\overline{\mathbb{R}}\right) \times \ldots \times F_d\left(\overline{\mathbb{R}}\right)$. Conversely, if C is a copula and $F_1, \ldots, F_d$ are univariate distribution functions, then function H defined above is a multivariate distribution function with margins $F_1, \ldots, F_d$.

Let X be the random vector behind the population $\mathcal{X}$, from which we extract a sample, but we are interested in the population $\mathcal{Y}$, governed by the random vector Y. We assume from now that X and Y share the copula C. This assumption is obviously valid when $\mathcal{X}$ is a Monte Carlo sample from $\mathcal{Y}$, and we argue in the following that it also holds when one population is a subset from the second one or when they are two different populations with similar characteristics. We formalize this idea by introducing the definition of a shared copula.

Définition 2 (Shared copula)**.** *The random vectors X and Y are said to share a copula C if there exists a copula C such that for any vector $(x_1, \ldots, x_d) \in \overline{\mathbb{R}}^d$, $F^X(x_1, \ldots, x_d) = C\left(F_1^X(x_1), \ldots, F_d^X(x_d)\right)$ and $F^Y(x_1, \ldots, x_d) = C\left(F_1^Y(x_1), \ldots, F_d^Y(x_d)\right)$.*

Since a copula is a multivariate distribution on [0,1]^d, the first step is to cast the observations as vectors in [0,1]^d. This can be easily achieved as it is standard in machine learning to apply some normalization procedure to the data (see for instance Larose and Larose (2014)), for instance by considering the empirical CDF (ECDF) of each of the feature X_i, $i = 1, \ldots, d$, defined as

$$\hat{F}_i(x) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{x_i \leq x}, \tag{3.3.2}$$

where $\mathbb{1}$ is the indicator function. The multivariate distribution underlying the normalized observations can be estimated using any of the previous considered methods.

Suppose now we wish to generate synthetic data of a target population from which we only have marginals information and that we have access to a sample of another, source, population sharing the structure of the target. Our population synthesis procedure is summarized in Algorithm 1.

Algorithme 1 Synthetic population generation

Step 1 Normalize the source population data using the ECDFs $\hat{F}_i(\cdot)$, $i = 1, \ldots, d$.
Step 2 Train the model on the normalized data to learn a copula C.
Step 3 Generate a synthetic population of vectors in $[0,1]^d$ by sampling from C.
Step 4 Transform any generated vector $\boldsymbol{u} = (u_1, \ldots, u_d)$ in a vector $\boldsymbol{y}$ in the target population as

$$\boldsymbol{y} = \left((F_1^Y)^{-1}(u_1), \ldots, (F_d^Y)^{-1}(u_d)\right),$$

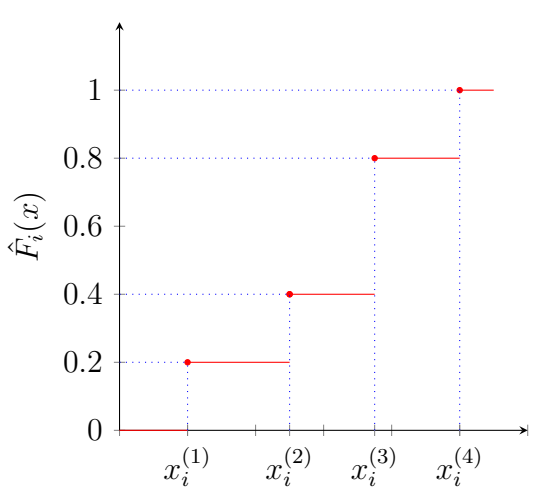
where $(F_i^Y)^{-1}(\cdot)$ is the pseudo-inverse distribution function of the i-th target marginal, defined as

$$(F_i^Y)^{-1}(u) = \min\left\{x_i : F_i^Y(x_i) \geq u\right\}.$$

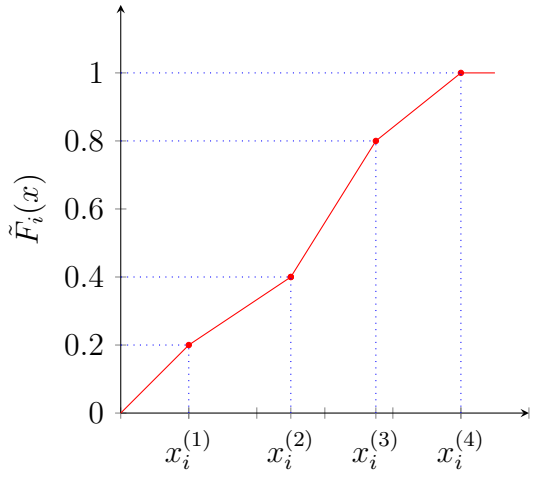
Algorithm 1 implicitly assumes that the marginal distribution functions of the source and target populations have the same ranges, an assumption often violated, especially when the marginals follow discrete distributions, which is especially the case when we use the ECDFs (3.3.2), as illustrated in Figure 3.2a. We address this issue by relaxing any discrete distribution function as a continuous distribution with the following heuristic. Consider the ECDF $\hat{F}_i(\cdot)$, with the range $\{x_i^{(1)}, \ldots, x_i^{(m_i)}\}$, and $x_i^{(p)} < x_i^{(q)}$ if $p < q$. We construct the relaxed ECDF $\tilde{F}_i(\cdot)$ as the continuous piecewise linear function obtained by considering the linear interpolation between and consecutive values $\hat{F}_i(x_i^k)$ and $\hat{F}_i(x_i^{k+1})$, $k \in (1,2,\ldots,m_i - 1)$, as depicted in Figure 3.2b. The relaxed marginals being now continuous, each one is uniformly distributed on [0,1]. We now extend the copula C to the domain $[0,1]^d$ by setting $C\left(\tilde{F}_1(x_1), \ldots, \tilde{F}_d(x_d)\right)$ as the linear interpolation, component by component, of $C\left(\hat{F}_1(x_1^{(k_1)}), \ldots, \hat{F}_d(x_d^{(k_d)})\right)$ and $C\left(\hat{F}_1(x_1^{(k_1+1)}), \ldots, \hat{F}_d(x_d^{(k_d+1)})\right)$, with $x_i^{(k_i)} \leq x_i \leq x_i^{(k_i+1)}$, setting $x_i^{(k_i+1)} = \infty$ (and thus $\hat{F}_d(x_i^{(k_i+1)}) = 1$) if $x_i^{(k)}$ corresponds to the greatest possible realisation of the i-th source marginal. Note that the copula built on the relaxed distribution functions still satisfies Sklar's theorem for the source marginals since it produces the same realizations on the Cartesian product of the range of the source marginals, but is now uniquely defined as any marginal distribution $\tilde{F}_i(\cdot)$, $i = 1, \ldots, d$ follows an uniform distribution on [0,1].

In short, we replace Step 3 in Algorithm 1 by

Step 3 Generate a synthetic population of vectors in $[0,1]^d$ by sampling from C extended to $\tilde{F}_i(\cdot)$, $i = 1, \ldots, d$.



(a) Empirical cumulative distribution function.



(b) Empirical cumulative distribution function with piecewise linear interpolations.

Fig. 3.2. Empirical cumulative distribution function and its relaxed continuous extension

3.4. Numerical experiments

To evaluate the methods described in the previous section, we generate synthetic populations at the state, county, and census tract levels. For that purpose, we use a sample of variables obtained from the American Community Survey (ACS), as well as their actual distributions (marginals), collected from the Decennial Census of Population Housing Data and the Internal Revenue Service (IRS). The ACS is a national demographic survey conducted every year by the U.S. Census Bureau that gathers sociodemographic and socioeconomic information at the household and individual level. This information is provided for geographical units called Public Use Micro Areas (PUMAs) that contain at least a hundred thousand individuals and do not overlap or nest within a single state. In our case, we use the 5-year ACS sample for the years 2012 to 2016 to ensure having a representative sample of the American population, as well as data from the 2010 census for the marginals. Table 3.1 below lists the nine selected variables, their names, level of aggregation, definitions, and associated values. With respect to the marginals, the decennial census data is updated every ten years at years ending with zeros and provides actual counts of the total number of individuals and households for several demographics and socioeconomic variables, including those of interest in our study.

In order to assess the quality and goodness of fit of the synthetic data we report two metrics: *the standardized root mean squared error* (SRMSE) and the count known as *sampling zeros* (SZ). Regarding the SRMSE, we compute it as in Sun and Erath (2015):

$$SRMSE = \sqrt{M \sum_{m_1=1}^{M_1} \cdots \sum_{m_d=1}^{M_d} (\pi_{m_1 \ldots m_d} - \hat{\pi}_{m_1 \ldots m_d})^2}, \tag{3.4.1}$$

Tableau 3.1. List of variables

Name	Definition	Level	Values
AGEP	Age of person	Individual	0 to 99 years
SEX	Gender of person	Individual	**1:** Male, **2:** Female
RAC1P	Race of person	Individual	**1:** White alone, **2:** Black or African American alone, **3:** American Indian alone, **4:** Alaska Native alone **5:** American Indian and Alaska Native and no other races,**6:** Asian alone, **7:** Native Hawaiian and Other Pacific Islander alone, **8:** Some other race alone, **9:** Two or more major race groups
ESR	Employment status	Individual	**1:** Civilian employed, at work, **2:** Civilian employed, with a job but not at work, **3:** Unemployed, **4:** Armed Forces, At Work, **5:** Armed Forces, With a Job but Not at Work, **6:** Not in Labor Force
HINCP	Household income (past 12 months)	Household	**1:** $1 to less $25k, **2:** $25k to less $50k, **3:** $50k to less $75k, **4:** $75k to less $100k, **5:** $100k to less $200k, **6:** $200k or more
HHT	Household/family type	Household	**1:** Married couple household, **2:** Male householder, no spouse present, **3:** Female householder, no spouse present, **4:** Male householder: Living alone, **5:** Male householder: Not living alone, **6:** Female householder: Living alone, **7:**Female householder: Not living alone
NP	Number of persons the household	Household	**1,6:** Number of persons in household **7:** Household of 7 persons or more
WIF	Workers in family during the past 12 months	Household	**0:** No workers, **1:** 1 worker, **2:** 2 workers, **3:** 3 or more workers in family
HUPAC	HH presence and age of children	Household	**1:** With children under 6 years only, **2:** With children 6 to 17 years only, **3:** With children under 6 years and 6 to 17 years, **4:** No children

where $\pi_{m_1 \ldots m_d}$ and $\hat{\pi}_{m_1 \ldots m_d}$ are the relative frequencies of a particular combination in the reference data and in the synthetic data, respectively and $M = \prod_{i=1}^{d} M_i$. SRMSE captures whether a combination of synthetic data appears in the actual data in similar proportions. A value of 0 means a perfect match while larger values evidence increasing distance between true and synthetic data. Moreover, we report the average SRMSE over subsets of n dimensions; when $n = 1$, the SRMSE measures the fitting of the marginals, while when $n > 1$ it measures the fitting of the multivariate dependencies.

SRMSE does not take into account the diversity of data. A synthetic combination might be desired even if it does not appear in the reference data. To assess the diversity of the produced synthetic data, we implement SZ, the count of the combinations of variables that are in the test set but not in the training set (Garrido et al., 2020). A SZ count of 0 means

that the generative model was unable to produce unseen realistic examples that do not appear in the training set, but do appear in the test set. Larger sampling zero values means that the generative model can output out-of-sample examples.

3.4.1. Population synthesis for the State of Maryland

This section is devoted to the task of generating synthetic data for a target population of which only a sub-sample and the marginal distributions are available. We compare BN, CTGAN, and TVAE to their copula counterparts, which means that we use them in the second and third steps of Algorithm 1 with the objective of determining whether the copula framework improves their performances in matching the marginal distributions of the target. These methods for which we applied the copula framework are respectively referred to as BN Copula, CTGAN Copula, and TVAE Copula. We also report the performances of IPF and that of an independent baseline created from the source population by sampling its variables independently.

In this section, we generate synthetic populations at the state and county levels. At the geographical level considered, the target population is the aggregation of available PUMAs, thus making it possible to assess the fitting of multivariate dependencies. To mimic how public agencies generally provide a subset from the whole population, we created a source population by randomly sampling 1% from the target. In addition to the source population, IPF, BN Copula, CTGAN Copula, and TVAE Copula were also given the empirical marginal distributions of all the target population's variables.

To evaluate the accuracy of the relationships between the variables of the generated data, we compute the average SRMSE between the synthetic and target populations over the possible combinations of n variables for n ranging from 1 to the total number of dimensions. We also report the number of sampled zeros to determine which models can generate unseen realistic examples.

On a technical note, we use the greedy search during the structural learning of the BN, then we generate values given the joint probability defined by the network via rejection sampling. For CTGAN and TVAE, we implement the network structures and hyperparameters suggested by Xu et al. (2019).

Table 3.2 presents SRMSE and SZ when the target population is the Sate of Maryland. It can be observed that IPF best captures marginal, bivariate, and trivariate distributions, but fails to capture higher order interactions between variables. In accordance with the findings of Sun and Erath (2015), unlike IPF, BN is able to capture complex relationships between variables. However, it is no better than the independent baseline at matching the target's marginals, which motivates the use of the copula framework. BN Copula indeed improves the SRMSE score for marginal distributions compared to BN. Overall, IPF performs best in

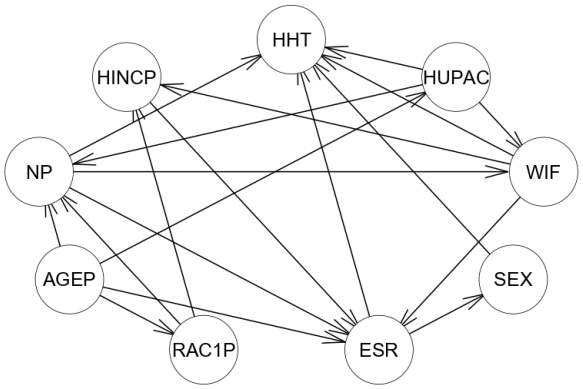


Fig. 3.3. Example of a Bayesian network found by the greedy search algorithm. It illustrates a factorization of the joint probability distribution of census variables.

small dimensions, closely followed by BN Copula, but, as dimensions increase, BN Copula significantly outperforms IPF.

	1	2	3	4	5	6	7	8	9	10	SZ
Ind	0.049	0.433	1.174	2.462	4.709	8.732	16.166	30.503	59.633	122.089	6937
CTGAN	0.399	0.866	1.630	2.976	5.438	10.067	19.007	36.761	73.087	149.743	23053
CTGAN Copula	0.362	0.789	1.490	2.723	4.965	9.154	17.199	33.136	65.856	135.591	24786
TVAE	0.425	1.099	2.249	4.271	7.899	14.515	26.770	49.906	94.666	183.920	18792
TVAE Copula	0.409	1.050	2.138	4.033	7.393	13.448	24.542	45.313	85.366	165.563	21230
BN	0.047	0.222	0.591	1.326	2.779	5.680	11.586	23.983	50.957	111.703	42447
BN Copula	0.009	0.183	0.546	**1.270**	**2.710**	**5.599**	**11.494**	**23.881**	**50.841**	**111.555**	41575
IPF	**0.001**	**0.129**	**0.496**	1.567	4.592	12.925	35.395	94.869	249.543	644.951	0

Tableau 3.2. State level SRMSE values between the target population and synthetic agents, column n represents the SRMSE averaged over the possible n-tuples variables. The last column lists sampling zero counts.

The county-level experiment, whose target is the aggregation of PUMAs in Anne Arundel County, shows very similar results. Table 3.3 presents the SRMSE and SZ metrics; BN Copula consistently succeeds in capturing the marginal distributions as well as the multivariate dependencies, outperforming IPF even in small dimensions.

Figures 3.4 and 3.5 plot the marginal distributions of the County and PUMA variables respectively from the State and County experiments. It can be observed that, compared to BN, BN Copula's counts are closer to those of the target.

In both experiments, the expected inability of IPF to generalize to new realistic out-of-sample examples is exposed. This is of great relevance since, although it may be a good

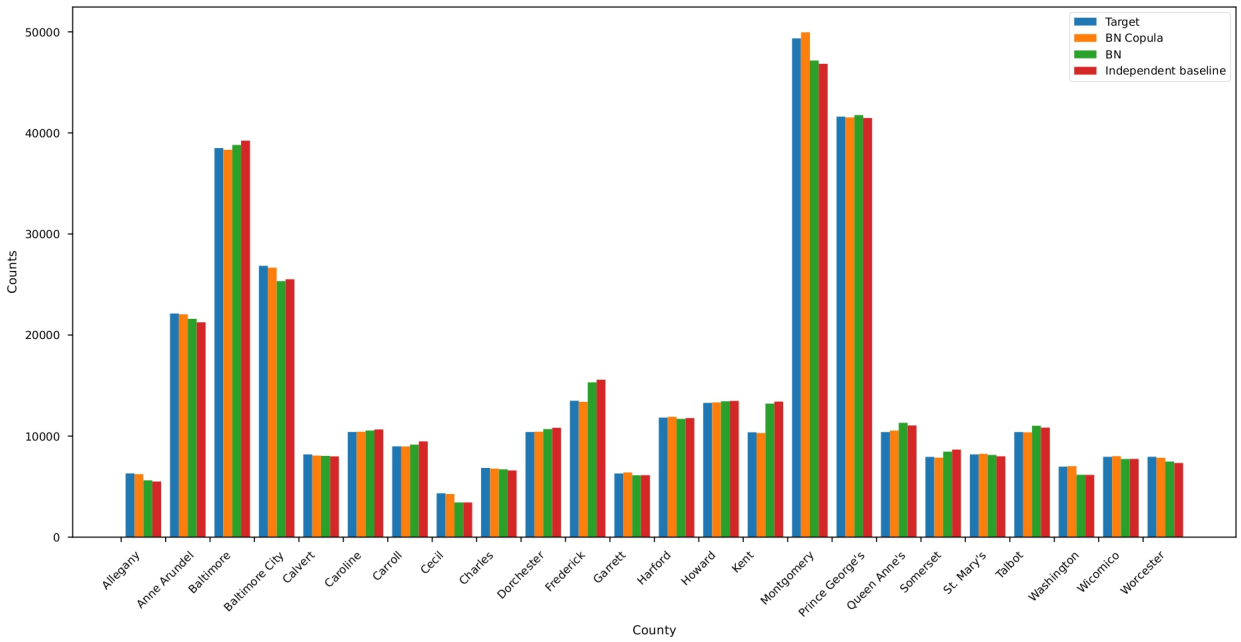


Fig. 3.4. Marginal fit of the county for the state level experiment

Ind	0.165	0.560	1.309	2.655	5.087	9.623	18.477	36.655	75.779	163.087	291
CTGAN	0.361	0.865	1.688	3.103	5.622	10.270	19.257	37.588	76.894	164.451	459
CTGAN Copula	0.268	0.688	1.421	2.725	5.102	9.578	18.368	36.478	75.530	162.715	480
TVAE	0.271	0.770	1.786	3.848	8.035	16.554	33.941	69.598	143.166	295.982	923
TVAE Copula	0.167	0.533	1.272	2.729	5.605	11.302	22.732	46.102	94.990	199.617	1216
BN	0.165	0.506	1.163	2.388	4.670	9.024	17.673	35.665	74.711	162.174	672
BN Copula	**0.015**	**0.303**	**0.899**	**2.042**	**4.213**	**8.431**	**16.928**	**34.771**	**73.691**	**161.081**	748
IPF	0.031	0.540	1.972	5.827	15.804	41.012	103.596	256.709	626.142	1505.718	0

Tableau 3.3. County level SRMSE values between the target population and synthetic agents, column n represents the SRMSE averaged over the possible n-tuples variables. The last column lists sampling zero counts.

procedure to capture the marginal distributions and the low-dimensional relationships, the synthetic populations it generates suffer from a significant lack of diversity.

It can also be noticed that all methods in conjunction with copulas perform better than their standalone counterparts. Moreover, a common trend among these methods is that the performance gaps observed narrow as the number of dimensions increases. We argue that this advantage of the copula framework is due to its ability to incorporate and match marginal distributions of the target, and that the effect of this advantage is less and less present as inter-dimensional relationships become more complex in size.

We should also mention that the proposed method heavily relies on the underlying model: its performance will follow the capacity of the generative model to learn the copula. CTGAN and TVAE poorly learn the population's distribution, which is why we can't expect

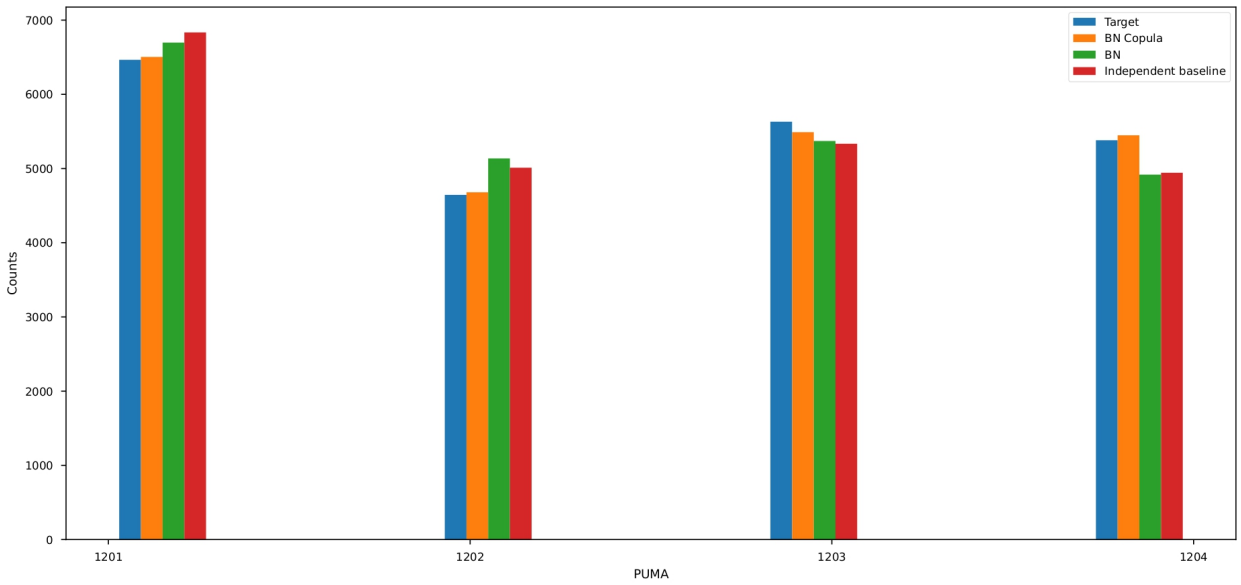


Fig. 3.5. Marginal fit of the PUMA for the county level experiment

their copula counterparts to perform well. Conversely, the proposed framework inherits the complexity and diversity of the underlying model when it is well designed for the given modeling task. According to these two state and county-level experiments, BN Copula is the most consistent model; like classical BN it captures high-dimensional relationships between variables and is able to sample zero cells, and also matches the marginal distributions of the target.

3.4.2. Population synthesis at the census tract level

We test our methodology at a smaller geographical level and we analyze the case when the source sample is a PUMA and the target population is a census tract inside that PUMA. In this experiment no sample of the target population is known. Therefore, classical ML frameworks are not applicable. However, we do know the marginal distributions of NP, RAC1P, SEX, HUPAC, and HHT at the census tract level from the decennial census 2010. We also estimate the marginal distribution for HINCP at the census tract level from the IRS data. We do so by counting the total number of tax returns by the different income categories, which included the totals of single returns, joint returns, head of family returns, and the number of dependents. We use these marginal distributions to estimate a BN Copula at PUMA level and we generate the synthetic population at census tract level assuming that the dependency structure can be maintained across the two geographical levels. We compare these results with those obtained with the modified IPF implemented in the PopGen software; this procedure uses two level of spatial resolutions, and in particular applies constraints at both PUMA and census tract levels. SRMSE cannot be computed in this case, and the two

methods are compared based on the marginals for the variables considered. Results relative to PUMA 1201 and census tract 702204 in Anne Arundel County are shown in Figure 3.6.

At this small geographical level, IPF is close to the source but fails to match the marginals of the target for most of the variables included in our analysis. We argue that IPF might face the zero-cell problem and is unable to create the observations that are in the target population (Guo and Bhat, 2007); when a demographic group of the target population is not represented in the source sample, its weights remain zero in the contingency table so that it cannot be sampled at the allocation stage. In contrast, the BN Copula method is able to produce a significant number of out-of-sample observations and fits almost perfectly the target's marginals. These results allow us to conclude that the proposed copula framework combined with BN is able to reproduce marginals at small area. This is relevant for producing statistics and estimates at small area level, which are more and more of interest to researchers and planning agencies.

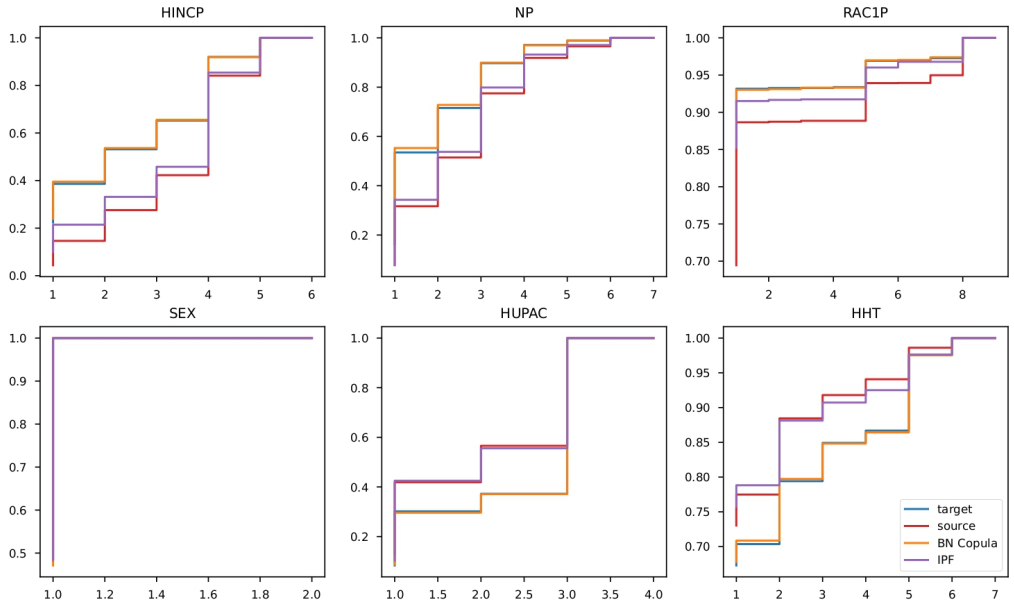


Fig. 3.6. Marginals fit for tract 702204 when source is PUMA 1201

3.4.3. Learning the dependencies from one PUMA to another

In previous experiments at the state, county, and census tract levels, the source population was obtained from a subregion of the target population. However, as stated in section 3.3, we claim that different populations with similar characteristics may share a copula even if the source is not subset of the target. In this experiment, we propose a spatial

transferability of a trained model that captures the dependencies of the selected variables from one PUMA to another. The concept of model transferability has been widely used in transportation forecasting modeling between different regions when data is unavailable or partially available, which reduces the time, cost, and effort of the data collection process Sikder et al. (2013); Bowman et al. (2014).

We tested our idea on several pairs of PUMAs and they all give similar results. Here we exhibit the results for a random pair of PUMAs selected to be the source and target populations.

The model transferability procedure cannot be executed on the classical ML methods, therefore we compare IPF and the copula-based frameworks since they both can incorporate the knowledge of the target's marginal distributions. According to the findings of our experiments at the state and county levels, BN seems more suited to the relatively low-dimensional synthetic population generation problem at hand, hence we report in Table 3.4 the performance of the copula framework only applied to the BN. The SRMSE values show that that IPF is slightly better than BN Copula at fitting the marginals for low-dimensional settings (up to five dimensions). However, as the dependencies grow, IPF again fails to capture the relationships while the SRMSE for the copula framework stays below the independent baseline. It should be stressed that IPF only replicates observations from the source and will always show null values in the sampling zeros column.

	1	2	3	4	5	6	7	8	9	SZ
Ind	0.179	0.634	1.601	3.540	7.513	15.971	34.577	76.500	172.372	37
BN Copula	**0.040**	0.281	0.901	2.342	5.654	**13.289**	**30.958**	**71.979**	**167.354**	182
IPF	0.042	**0.218**	**0.722**	**2.033**	**5.341**	13.577	33.818	82.975	200.853	0

Tableau 3.4. SRMSE values between the target population and synthetic agents for the PUMA to PUMA experiment, column n represents the SRMSE averaged over the possible n-tuples variables. The last column lists sampling zero counts.

3.5. Conclusion

In this paper, we introduce a framework that integrates the theory of copulas into traditional machine learning generative methods to capture both joint and marginal distributions in synthetic population generation. We also explore the possibility to transfer the copula trained on one region and learn about the joint distribution of the variables in a different one. The classical data normalization relying on empirical distribution functions of the features allows us to consider the observations as realizations of an unknown copula, capturing the dependencies between their components, which can be learned using standard techniques. We propose a heuristic approach to overcome the issues related to discrete distributions, and propose a simple algorithm to generate a synthetic population with different marginals but

the same dependencies structure. Since the geographical level is a crucial factor for population synthesis and can influence the performance of different deep generative methods, we conduct comparison experiments at the state, county, and census tract levels. We apply the proposed copula framework to each of the selected ML generative methods: BN, CTGAN, and TVAE, which are suitable for the size of our problem and the nature of our data. Both accuracy and diversity are important criteria to evaluate the performance of different synthetic population generation methods, so we use SRMSE to evaluate the accuracy of the synthetic population and sampling zeros to evaluate the diversity. We compare the performance of these methods and their copula counterparts to IPF and a baseline population obtained from the ACS sample source by bootstrapping its dimensions independently.

The results of the experiments at the three geographical levels considered show that the copula framework improves the performance of all the ML methods under study. Also, the BN Copula always outperforms the BN method at both state and county level based on the SRMSE metric and the ability to recover the marginal distributions of the variables that characterize the population. At census tract level, where no sample is available – which makes conventional ML techniques not applicable, the IPF with double geographical constraints fails to recover the target values for most of the variables considered, while the BN Copula follows the expected patterns. The proposed spatial transferability, which is again not feasible with standard ML techniques, shows that at PUMA level the BN Copula framework always produces results that are superior to the independent case, and that are close, or even better, to the IPF (up to five dimensions, or more than five, respectively); while, at the same time, producing observations that are not in the original sample. Nevertheless, it is worth mentioning that our study included only discrete variables and not continuous variables. Hence, we have not been able to explore the performance of the proposed copula framework in these cases. This is an extension of this work that we plan to carry out in the future. Namely, to explore more ML generative methods and different data types to prove our conclusions in different methodological and information contexts. Finally, this study opens up to the opportunity to combine data from multiple sources (e.g. disaggregated survey data and aggregated administrative data) to overcome the limited sample size problem for small areas. This is especially important when producing population statistics and travel indicator estimates at small geographical level which are critical for policy analysis in governmental agencies.

Acknowledgments

The work of Fabian Bastin is supported by the Natural Sciences and Engineering Research Council of Canada [Discovery Grant 2022-04400]. Pascal Jutras-Dubé was furthermore supported by a graduate grant under the NSERC CREATE Program on Machine Learning in Quantitative Finance and Business Analytics (Fin-ML).

Références bibliographiques

Theo Arentze, Harry J. P. Timmermans, and Frank Hofman. Creating synthetic household populations: Problems and approach. *Transportation Research Record*, 2014:85–91, 2007.

Theo A. Arentze and Harry J.P. Timmermans. A learning based transportation oriented simulation system. *Transportation Research Part B*, 38(7):613–633, 2004.

Joshua Auld, Abolfazl Mohammadian, and Kermit Wies. Population synthesis with subregion-level control variable aggregation. *Journal of Transportation Engineering*, 135 (9):632–639, 2009.

Joshua Auld, Mahmoud Javanmardi, and Abolfazl Mohammadian. Integration of activity scheduling and traffic assignment in ADAPTS activity-based model. In *TRB 91st Annual Meeting Compendium of Papers DVD*, number 12-4225, Washington DC, United States, January 2012. Transportation Research Board.

Athanassios Avramidis, Nabil Channouf, and Pierre L'Ecuyer. Efficient correlation matching for fitting discrete multivariate distributions with arbitrary marginals and normal-copula dependence. *INFORMS Journal on Computing*, 21:88–106, 2009.

Johan Barthelemy and Philippe L. Toint. Synthetic population generation without a sample. *Transportation Science*, 47(2):266–279, 2013.

Chandra R. Bhat and Naveen Eluru. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B*, 43(7): 749–765, 2009.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New-York, NY, USA, 2006.

Guillem Boquet, Antoni Morell, Javier Serrano, and Jose Lopez Vicario. A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection. *Transportation Research Part C*, 115: 102622, 2020.

Edoardo Borgomeo, Georg Pflug, Jim W. Hall, and Stefan Hochrainer-Stigler. Assessing water resource system vulnerability to unprecedented hydrological drought using copulas to characterize drought duration and deficit. *Water Resources Research*, 51(11):8927–8948, 2015.

Kathryn Born, Shamsunnahar Yasmin, Daehyun You, Naveen Eluru, Chandra R. Bhat, and Ram M. Pendyala. Joint model of weekend discretionary activity participation and episode duration. *Transportation Research Record*, 2413(1):34–44, 2014.

Stanislav S. Borysov and Jeppe Rich. Introducing synthetic pseudo panels: application to transport behaviour dynamics. *Transportation*, 48(5):2493–2520, 2021.

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*, 106:73–97, 2019.

John L. Bowman, Mark Bradley, Joe Castiglione, and Supin L. Yoder. Making advanced travel forecasting models affordable through model transferability. In *Transportation Research Board 93rd Annual Meeting*, January 2014.

Mark Bradley, John L. Bowman, and Bruce Griesenbeck. Sacsim: An applied activity-based model system with fine-level spatial and temporal resolution. *Journal of Choice Modelling*, 3(1):5–31, 2010.

Daniele Casati, Kirill Müller, Pieter J Fourie, Alexander Erath, and Kay W Axhausen. Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record*, 2493(1):107–116, 2015.

Umberto Cherubini, Elisa Luciano, and Walter Vecchiato. *Copula Methods in Finance*. John Wiley & Sons, Chichester, United Kingdom, 2004.

Gregory F. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2):393–405, 1990.

W. Edwards Deming and Frederick F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

Gérald Duguay, Woo Jung, and Daniel L. McFadden. SYNSAM: A methodology for synthesizing household transportation survey data. Working paper 7618, Institute of Transportation Studies, University of California, Berkeley, CA, USA, 1976.

Naveen Eluru, Abdul Rawoof Pinjari, Jessica Y. Guo, Ipek Nese Sener, Sivaramakrishnan Srinivasan, Rachel B. Copperman, and Chandra R. Bhat. Population updating system structures and models embedded in the comprehensive econometric microsimulator for urban systems. *Transportation Research Record*, 2076:171–182, 2008.

Bilal Farooq, Michel Bierlaire, Ricardo Hurtubia, and Gunnar Flötteröd. Simulation based population synthesis. *Transportation Research Part B*, 58:243–263, 2013.

Sergio Garrido, Stanislav S. Borysov, Francisco C. Pereira, and Jeppe Rich. Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C*, 120:102787, 2020.

Christian Genest and Anne-Catherine Favre. Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4):347–368, 2007.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Lukas Günthermann, Ivor Simpson, and Daniel Roggen. Smartphone location identification and transport mode recognition using an ensemble of generative adversarial networks. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp-ISWC '20, pages 311—316, New York, NY, USA, September 2020. Association for Computing Machinery.

Jessica Guo and Chandra Bhat. Population synthesis for microsimulating travel behavior. *Transportation Research Record*, 2014:92–101, 2007.

David Heckerman, Dan Geiger, and David M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):194–243, 1995.

Zengyi Huang and Paul Williamson. A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata. Working Paper 2001/2, University of Liverpool, October 2001.

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*, 130:103291, 2021.

Amel Jaoua, Pierre L'Ecuyer, and Louis Delorme. Call-type dependence in multiskill call centers. *SIMULATION*, 89(6):722–734, 2013.

Harry Joe. *Multivariate Models and Dependence Concepts*. Springer, New York, NY, USA, 1997.

Harry Joe. *Dependence modeling with copulas*. CRC Press, Boca Raton, FL, USA, 2015.

Shih-Chieh Kao, Hoe Kyoung Kim, Cheng Liu, Xiaohui Cui, and Budhendra L. Bhaduri. Dependence-preserving approach to synthesizing household characteristics. *Transportation Research Record*, 2302:192–200, 2012.

Kartik Kaushik, Cinzia Cirillo, and Fabian Bastin. On modelling human population characteristics with copulas. *Procedia Computer Science*, 151:210–217, 2019. The 10th International Conference on Ambient Systems, Networks and Technologies (ANT 2019) / The 2nd International Conference on Emerging Data and Industry 4.0 (EDI40 2019) / Affiliated Workshops.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, Banff, AB, Canada, April 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*, 2563:40–50, 2016.

Wai Lam and Fahiem Bacchus. Learning bayesian belief networks: An approach based on the mdl principle. *Computational Intelligence*, 10(3):269–293, 1994.

Daniel T. Larose and Chantal D. Larose. *Data Preprocessing*, chapter 2, pages 16–50. John Wiley & Sons, Hoboken, NJ, USA, 2014.

Kirill Müller and Kay W. Axhausen. Population synthesis for microsimulation: State of the art. In *TRB 90th Annual Meeting Compendium of Papers DVD*, number 11-1789, Washington DC, United States, January 2011. Transportation Research Board.

Roger B Nelsen. *An Introduction to Copulas*. Springer, New York, NY, USA, second edition, 2006.

Ostap Okhrin, Alexander Ristig, and Ya-Fei Xu. *Copulae in High Dimensions: An Introduction*, chapter 13, pages 247–277. Springer, Berlin, Germany, third edition, 2017.

Boris Oreshkin, Nazim Réegnard, and Pierre L'Ecuyer. Rate-based daily arrival process models with application to call centers. *Operations Research*, 64, 2016.

Abdul Rawoof Pinjari, Chandra R. Bhat, and David A. Hensher. Residential self-selection effects in an activity time-use behavior model. *Transportation Research Part B*, 43(7): 729–748, 2009.

David R. Pritchard and Eric J. Miller. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, 39(3):685–704, 2012.

Tejsingh A. Rana, Sujan Sikder, and Abdul Rawoof Pinjari. Copula-based method for addressing endogeneity in models of severity of traffic crash injuries: Application to two-vehicle crashes. *Transportation Research Record*, 2147:75–87, 2010.

Ismaïl Saadi, Ahmed Mustafa, Jacques Teller, and Mario Cools. Forecasting travel behavior using markov chains-based approaches. *Transportation Research Part C*, 69:402–417, 2016a.

Ismaïl Saadi, Ahmed Mustafa, Jacques Teller, Bilal Farooq, and Mario Cools. Hidden markov model-based population synthesis. *Transportation Research Part B*, 90:1–21, 2016b.

Paul Salvini and Eric J. Miller. ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems. *Networks and spatial economics*, 5(2):217–234, 2005.

Sujan Sikder, Abdul Rawoof Pinjari, Sivaramakrishnan Srinivasan, and Nowrouzian Roosbeh. Spatial transferability of travel forecasting models: a review and synthesis. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, 5(2): 104–128, 2013.

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

Laron Smith, Richard Beckman, Keith Baggerly, Doug Anson, and Michael Williams. TRANSIMS: Transportation analysis and simulation system. techreport LA-UR-95-1641, Los Alamos National Laboratory, Los Alamos, NM, USA, July 1995.

Lijun Sun and Alexander Erath. A Bayesian network approach for population synthesis. *Transportation Research Part C*, 61:49–62, 2015.

Lijun Sun, Alexander Erath, and Ming Cai. A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B*, 114:199–212, 2018.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional GAN. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, and Emily B. Fox, editors, *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 7335–7345, Red Hook, NY, USA, 2019. Curran Associates Inc.

Ali Yazdizadeh, Zachary Patterson, and Bilal Farooq. Semi-supervised GANs to infer travel modes in GPS trajectories. *Journal of Big Data Analytics in Transportation*, 3(3):201–211, 2021.

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 *TRB 88th Annual Meeting Compendium of Papers DVD*, number 09-2096, Washington DC, United States, jan 2009. Transportation Research Board.

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.

Danqing Zhang, Junyu Cao, Sid Feygin, Dounan Tang, Zuo-Jun(Max) Shen, and Alexei Pozdnoukhov. Connected population synthesis for transportation simulation. *Transportation Research Part C*, 103:1–16, 2019.

Dominik Ziemke, Johan W. Joubert, and Kai Nagel. Accessibility in a post-apartheid city: Comparison of two approaches for accessibility computations. *Networks and Spatial Economics*, 18(2):241–271, 2018.

Chapitre 4

Définitions de confidentialité

Cette section discute les garanties théoriques des principales approches modernes en matière de publication de données qui resepecte la confidentialité. On y définit le k-anonymat qui adresse le risque de ré-identification au moyen de généralisation et de suppression de champs quasi-identifiants, les données synthétiques qui remplacent complètement les données sensibles et la confidentialité différentielle qui adresse le risque de divulgation de présence en s'assurant au moyen de randomisation que quelqu'un qui observe les données publiées ne peut découvrir si une personne est présente ou non dans le jeu de données sensibles.

4.1. k-anonymat

Le concept du k-anonymat a été introduit pour la première fois par Sweeney (2002) pour garantir que les individus d'un jeu de données ne peuvent pas être ré-identifiés en reliant les données publiées à des données auxiliaires.

Définition 4.1.1 (k-anonymat)**.** *Un jeu de données est dit k-anonyme ou posséder la propriété du k-anonymat si chacun de ses enregistrements partage les valeurs d'au moins $k-1$ autres enregistrements.*

En d'autres mots, un jeu de données k-anonyme est une partition dont les classes d'équivalence sont de tailles au moins k. La figure 4.1 présente un exemple de données 2-anonymes.

Les champs quasi-identifiants (QI) sont ceux qui combinés peuvent ré-identifer un individu. S'ils sont connus, il suffit qu'eux seuls adhèrent au k-anonymat pour prévenir le risque de ré-identification. Cependant, un nombre exponentiel de combinaisons de dimensions peut potentiellement servir aux attaques de liaisons (Aggarwal, 2005).

Par ailleurs, remarquons que le risque de ré-identifcation est majoré par $1/k$ parce qu'un observateur peut au mieux ré-identifier la classe d'équivalence dans laquelle se trouve l'individu qu'il cherche, mais que cette garantie ne tient pas pour des données corrélées; un utilisateur qui est représenté par plusieurs enregistrements court un risque de ré-identification potentiellement plus élevé.

Âge	Éducation	Occupation	Ville	Revenu
$20 < \text{Âge} < 30$	Baccalauréat	Ingénieur	Montréal	$70k < \text{Revenu} < 80k$
$30 < \text{Âge} < 40$	Doctorat	Chercheur postdoctoral	Vancouver	$60k < \text{Revenu} < 70k$
$20 < \text{Âge} < 30$	Baccalauréat	Ingénieur	Montréal	$80k < \text{Revenu} < 90k$
$30 < \text{Âge} < 40$	Doctorat	Chercheur postdoctoral	Vancouver	$70k < \text{Revenu} < 80k$
$30 < \text{Âge} < 40$	Doctorat	Chercheur postdoctoral	Vancouver	$70k < \text{Revenu} < 80k$
$\text{Âge} < 20$	Secondaire	Étudiant	Montréal	$\text{Revenu} < 20k$
$\text{Âge} < 20$	Secondaire	Étudiant	Montréal	$\text{Revenu} < 20k$

Tableau 4.1. Jeu de données dont les champs "Âge", "Éducation", "Occupation" et "Ville" adhèrent au 2-anonymat.

En général, les données brutes ne satisfont pas la propriété du k-anonymat et elle est atteinte au moyen de généralisation et de suppression. La généralisation consiste à remplacer ou recoder une valeur par une valeur sémantiquement cohérente mais moins spécifique tandis que la suppression consiste à ne pas publier une valeur.

L'algorithme Mondrian[1] (LeFevre *et al.*, 2006) partitionne les données en utilisant les arbres k-d (Bentley, 1975). Chaque itération sépare les données en deux groupes. Dans un premier temps, l'algorithme détermine une dimension selon laquelle séparer les données. Une heuristique populaire est de choisir celle ayant le plus grand diamètre normalisé. Pour les variables catégorielles, le diamètre est défini comme le nombre de catégories dans un groupe. Pour les variables numériques, il est défini comme la différence entre les maximum et minimum. Le diamètre dans un groupe est normalisé par le diamètre du jeu de données entier. Ensuite, l'algorithme sélectionne la valeur de séparation comme la médiane qui est un bon candidat pour les variables numériques. Pour les variables catégorielles, il s'agit de placer autant de catégories dans les deux groupes. Ce processus est répété récursivement dans les groupes ainsi générés jusqu'à ce qu'il ne soit plus possible de séparer les données en des groupes de tailles supérieures ou égales à k. À la toute dernière étape, les champs QIs sont généralisés au sein des groupes. Le pseudo code 2 résume les étapes de l'algorithme Mondrian pour sélectionner les classes d'équivalences.

4.2. Données synthétiques

Plutôt que de modifier directement les données sensibles pour les publier, elles peuvent être remplacées par des données synthétiques. Avec la distribution générative en main, il pourrait être possible d'échantillonner de fausses données qui respectent les propriétés statistiques de la population véritable. En plus de supprimer les individus réels de la publication de données, cette approche permet de générer un nombre arbitraire de données.

[1]C'est au peintre Piet Mondrian que l'algorithme doit son nom, la partition générée de données bidimensionnelles rappelant ses peintures.

Algorithme 2 Mondrian

```
classes ← [ ]
ANONYMISER(classe)
    if aucune coupe n'est permise then
        classes.AJOUTER(classe)
    end if
    dim ← CHOISIRDIM( )
    valeurSplit ← MÉDIANE(partition.dim)
    classeGauche ← {x ∈ classe : x.dim < valeurSplit}
    classeDroite ← {x ∈ classe : x.dim ≥ valeurSplit}
return ANONYMISER(classeGauche) ∪ ANONYMISER(classeDroite)
end
```

Les réseaux antagonistes génératifs (GAN, de l'anglais *generative adversarial network*) (Goodfellow *et al.*, 2014) sont une classe d'algorithmes d'apprentissage automatique non supervisés dans lesquels un modèle génératif est opposé à un modèle discriminatif qui apprend à déterminer si un échantillon provient de la distribution du modèle ou de la distribution des données.

Le générateur G_θ est un réseau de neurones dont les paramètres sont θ. Sa fonction est d'approximer la distribution des données p_X en génèrant des échantillons synthétiques $G_\theta(Z)$ à partir d'une distribution latente $Z \sim p_Z$. Le discriminateur D_ϕ est un réseau de neurones paramétré par ϕ dont la sortie scalaire $D_\phi(x)$ représente la probabilité qu'un exemple x vienne des vraies données plutôt que de la distribution du générateur. D_ϕ est entraîné pour maximiser la probabilité d'assigner la bonne classe aux exemples d'entraînement et aux échantillons de G_θ. En revanche, G_θ est entraîné pour minimiser la probabilité que D_ϕ attribue la bonne classe à sa sortie

$$\min_\theta \max_\phi \mathbb{E}_{X\sim p_X}[\log D_\phi(X)] + \mathbb{E}_{Z\sim p_Z}[\log(1 - D_\phi(G_\theta(Z)))].$$

En pratique, il semble que le terme $\log(1 - D_\phi(G_\theta(Z)))$ puisse saturer et ne pas fournir un signal d'apprentissage suffisant lors de l'entraînement du générateur par l'algorithme de descente du gradient stochastique (SGD, de l'anglais *stochastic gradient descent*). On optimise plutôt G_θ pour maximiser $\log D_\phi(G_\theta(Z))$

$$\max_\phi \mathbb{E}_{X\sim p_X}[\log D_\phi(X)] + \mathbb{E}_{Z\sim p_Z}[\log(1 - D_\phi(G_\theta(Z)))]$$
$$\max_\theta \mathbb{E}_{Z\sim p_Z}[\log D_\phi(G_\theta(Z))].$$

Les GAN ont atteint des performances impressionnantes dans la génération d'images synthétiques de haute qualité, mais ils ne sont pas explicitement conçus pour gérer des données tabulaires. Un réseau antagoniste génératif tabulaire conditionnel (CTGAN, de l'anglais *conditional tabular generative adversarial network*)(Xu *et al.*, 2019b), est une adaptation

des GANs conçu pour modéliser la distribution de probabilité des lignes dans les données tabulaires, qui contiennent généralement un mélange de colonnes discrètes et continues.

Typiquement les distributions des colonnes continues de données tabulaires possèdent plusieurs modes et ne ressemblent pas à des distributions gaussiennes. La transformation min-max généralement utilisée pour normaliser les données conjointement employée avec une non-linéarité tanh risque d'occasionner la saturation du gradient lors de la procédure d'optimisation par SGD. CTGAN utilise la normalisation spécifique au mode pour traiter les colonnes continues avec des distributions compliquées.

Pour une colonne continue C_i, la normalisation spécifique au mode estime d'abord le nombre de modes m_i avec un modèle de mélange gaussien variationnel (VGM, de l'anglais *variational Gaussian mixture*). Pour chaque mode η_k, $k = 1, \ldots, m_i$ le VGM apprend une densité gaussienne $\mu_k \mathcal{N}(\eta_k, \sigma_k)$ dont σ_k^2 et μ_k sont respectivement la variance et le poids. Ensuite, pour chaque valeur $c_{i,j}$ de C_i, un mode est échantillonné selon les probabilités qu'elle appartienne à chaque mode. Finalement $c_{i,j}$ est remplacée par un vecteur one-hot indiquant le mode échantillonné η_k et par une valeur scalaire $\alpha_{i,j} = \frac{c_{i,j} - \eta_k}{4\sigma_k}$.

Quant aux colonnes discrètes, elles sont souvent déséquilibrées de sorte que les catégories minoritaires risquent d'être sous-représentées lorsque les données sont échantillonnées aléatoirement lors du processus d'apprentissage. CTGAN fait usage d'un générateur conditionnel pour inciter une variable discrète D_{i^*} à prendre une valeur particulière $k^* \in \{1, \ldots, |D_{i^*}|\}$. Chaque colonne D_i est représentée par un encodage one-hot d_i et pour chacune d'elles la condition est exprimée par un vecteur m_i dont les composantes indicées par $k = 1, \ldots, |D_i|$ sont

$$m_i^{(k)} = \begin{cases} 1 & \text{si } i = i^* \text{ et } k = k^* \\ 0 & \text{sinon.} \end{cases}$$

Pour inciter le générateur à respecter la condition $D_{i^*} = k^*$, un terme d'entropie croisée pour chaque colonne discrète D_i est ajouté à la fonction de coût entre le vecteur one-hot $\hat{d}_i$ généré et le vecteur de condition m_i. Finalement, pour choisir la condition, CTGAN sélectionne d'abord au hasard une colonne discrète, puis construit une fonction de masse de probabilité sur la plage de valeurs de cette colonne et échantillonne une valeur selon cette fonction de masse.

4.3. Confidentialité différentielle

Une approche naturelle pour définir la confidentialité dans un contexte d'analyse de données consiste à exiger que l'analyste n'en sache pas plus sur un individu dans l'ensemble de données une fois l'analyse terminée qu'il n'en savait avant qu'elle ne commence. C'est-à-dire que la publication de données qui respecte la confidentialité devrait enseigner de l'information utile sur une population sans rien enseigner au sujet d'un individu présent dans les données.

Mais qu'est-ce que rien apprendre au sujet d'un individu? L'exemple du fumeur de Dwork et Roth (2014) nous en instille l'intuition. Supposons qu'une base de données médicales nous apprend que fumer cause le cancer, ce qui affecte l'opinion d'une compagnie d'assurance par rapport aux frais médicaux à long terme d'un fumeur. A-t-on compromis la confidentialité du fumeur? Certes le fumeur est affecté par l'étude, mais l'impact est le même qu'il y ait ou non participé. Il s'agit de nuancer entre le pouvoir prédictif des données publiées et la divulgation d'attributs d'un individu ré-identifié dans les données. La confidentialté différentielle (DP) (Dwork *et al.*, 2006; Dwork, 2011; Dwork et Roth, 2014) adresse cette distinction en garantissant au moyen de randomisation que les mêmes conclusions peuvent être tirées qu'un individu soit ou non dans l'ensemble de données.

Définition 4.3.1 (confidentialité différentielle)**.** *Un mécanisme aléatoire $\mathcal{M} : \mathcal{D} \to \mathcal{I}$ de domaine $\mathcal{D}$ et d'image $\mathcal{I}$ est ϵ-DP si pour tout jeux de données adjacents $X, Y \in \mathcal{D}$ et pour tout $S \subseteq \mathcal{I}$ on a*

$$P[\mathcal{M}(X) \in S] \leq \exp(\epsilon) P[\mathcal{M}(Y) \in S].$$

La probabilité est prise sur la randomisation du mécanisme $\mathcal{M}$. On dit que $X, Y \in \mathcal{D}$ sont adjacents lorsqu'ils diffèrent par au plus un enregistrement. L'ensemble S est un évènement de sortie de $\mathcal{M}$ comme une réponse à une requête ou un jeu de données synthétiques.

La confidentialité différentielle garantit que toute séquence de sortie est presque également susceptible de se produire indépendamment de la présence de tout individu et cette différence en probabilité est capturée par ϵ, le budget de confidentialité. En effet, lorsque ϵ est positif et petit, on a que

$$\frac{P[\mathcal{M}(X) \in S]}{P[\mathcal{M}(Y) \in S]} \leq \exp(\epsilon) \approx 1 + \epsilon.$$

Une alternative à la définition 4.3.1 permet qu'un évènement de sortie $\xi \sim \mathcal{M}$ ait lieu avec faible probabilité δ si un individu est présent dans un jeu de données mais qu'il n'arrive presque jamais autrement; en d'autres mots qu'il y ait une faible chance que $\mathcal{M}$ ne soit pas ϵ-DP (lemme 3.17 de Dwork et Roth (2014)).

Définition 4.3.2 (confidentialité différentielle approchée)**.** *Un mécanisme aléatoire $\mathcal{M} : \mathcal{D} \to \mathcal{I}$ de domaine $\mathcal{D}$ et d'image $\mathcal{I}$ est (ϵ,δ)-DP si pour tout jeux de données adjacents $X, Y \in \mathcal{D}$ et pour tout $S \subseteq \mathcal{I}$ on a*

$$P[\mathcal{M}(X) \in S] \leq \exp(\epsilon) P[\mathcal{M}(Y) \in S] + \delta.$$

Si un mécanisme $\mathcal{M}$ est (ϵ,δ)-DP, alors pour chaque paire de jeux de données adjacents X et Y, il est estrêmement rare que la sortie $\mathcal{M}(X)$ soit beaucoup plus (ou beaucoup moins) susceptible de se produire si le jeu de données est X que s'il est Y. C'est-à-dire qu'il est possible que pour une sortie $\xi \sim \mathcal{M}(X)$, on soit capable de trouver des jeux de données adjacents X et Y tels que la masse de ξ dans la distribution $\mathcal{M}(X)$ est considérablement plus large que dans la distribution $\mathcal{M}(Y)$ ou vice versa. L'éventualité indésirable pour

laquelle cette sortie révélatrice advient suit une loi binomiale de paramètres (n, δ) où n est le nombre d'individus dans les données. Conséquemment, on peut s'attendre à révéler l'information de $n\delta$ individus en moyenne. Par exemple, si $\delta = \frac{1}{100n}$, alors on peut assurer que ce mauvais scénario n'arrive pas avec une probabilité de 99%. En particulier, les valeurs $\delta \geq \frac{1}{n}$ sont dangereuses parce qu'elles permettent de publier les informations complètes de certains participants. On veut donc préférablement que $\delta \ll \frac{1}{n}$.

La définition de la confidentialité différentielle lui confère quelques garanties théoriques avantageuses. Pour étudier les propriétés de la confidentialité différentielle, il est convenable de définir la notion de perte de confidentialité.

Définition 4.3.3 (perte de confidentialité)**.** *Pour des jeux de données adjacents X et Y, un mécanisme $\mathcal{M}$ et un évènement de sortie ξ, la perte de confidentialité est la variable aléatoire*

$$\mathcal{L}_{\mathcal{M}(X)\|\mathcal{M}(Y)}(\xi) = \ln\left(\frac{P[\mathcal{M}(X) = \xi]}{P[\mathcal{M}(Y) = \xi]}\right).$$

La confidentialité différentielle est immunisée face au post-traitement: un observateur ne peut pas appliquer une fonction à la sortie du mécanisme de confidentialité dans l'espoir d'en découvrir d'avantage au sujet des données sensibles.

Proposition 4.3.4 (post-traitement)**.** *Soit $\mathcal{M} : \mathcal{D} \to \mathcal{I}$ un mécanisme aléatoire qui satisfait la (ϵ,δ)-DP et $f : \mathcal{I} \to \mathcal{I}'$ une fonction arbitraire. On a que $f \circ \mathcal{M} : \mathcal{D} \to \mathcal{I}'$ est (ϵ,δ)-DP.*

Démonstration. Soit X,$Y \in \mathcal{D}$ des jeux de données adjacents et soit $S \subseteq \mathcal{I}'$. Soit aussi $T = \{s \in \mathcal{I} : f(s) \in S\}$. On a

$$\begin{aligned} P[f(\mathcal{M}(X)) \in S] &= P[\mathcal{M}(X) \in T] \\ &\leq \exp(\epsilon) P[\mathcal{M}(Y) \in T] + \delta \\ &= \exp(\epsilon) P[f(\mathcal{M}(Y)) \in S] + \delta. \end{aligned}$$

□

En vertu de la propriété de post-traitement, une publication de données générée par un mécanisme DP est immunisée face aux attaques par liaison: le budget de confidentialité est indépendent des informations auxiliaires auxquelles un observateur pourrait avoir accès, ce qui réduit le risque de ré-identification.

D'autre part, la confidentialité différentielle tient pour des enregistrements corrélés.

Proposition 4.3.5 (confidentialité différentielle de groupe)**.** *Soit X et Y des jeux de données qui diffèrent par au plus k enregistrements et un mécanisme de confidentialité $\mathcal{M}$ qui est ϵ-DP, alors*

$$P[\mathcal{M}(X) \in S] \leq \exp(k\epsilon) P[\mathcal{M}(Y) \in S].$$

Démonstration. Le cas $k = 1$ est vrai par hypothèse. Par induction, posons que $k = n + 1$ et notons X' une base de données adjacente à Y et qui diffère de X par au plus n

enregistrements. X' existe parce que Y et X diffèrent par au plus $n+1$ enregistrements. On a que

$$\begin{aligned} P[\mathcal{M}(X) \in S] &\leq \exp(n\epsilon) P[\mathcal{M}(X') \in S] \\ &\leq \exp((n+1)\epsilon) P[\mathcal{M}(Y') \in S]. \end{aligned}$$

□

La propriété la plus pratique est probablement celle de la composition parce qu'elle permet la conception de mécanismes DP complexes.

Proposition 4.3.6 (composition de base)**.** *Soit les mécanismes indépendents $\mathcal{M}_1$ et $\mathcal{M}_2$ respectivement ϵ_1-DP et ϵ_2-DP. Le mécanisme $\mathcal{M}(\cdot) = (\mathcal{M}_1(\cdot), \mathcal{M}_2(\cdot))$ est $(\epsilon_1 + \epsilon_2)$-DP.*

Démonstration. Soit X et Y des jeux de données adjacents et (s_1, s_2) une sortie quelconque. On a

$$\frac{P[\mathcal{M}(X) = (s_1, s_2)]}{P[\mathcal{M}(Y) = (s_1, s_2)]} = \frac{P[\mathcal{M}_1(X) = s_1] P[\mathcal{M}_2(X) = s_2]}{P[\mathcal{M}_1(Y) = s_1] P[\mathcal{M}_2(Y) = s_2]} \leq \exp(\epsilon_1) \exp(\epsilon_2) = \exp(\epsilon_1 + \epsilon_2)$$

□

Une manière courante d'approximer une fonction déterministe par un mécanisme qui satisfait la confidentialité différentielle consiste à ajouter du bruit calibré par la sensibilité de cette fonction. Pour publier des données respectant la confidentialité différentielle, on pourrait appliquer du bruit directement sur les données ou sur les paramètres internes d'un modèle génératif, mais dans ces cas, caractériser la sensibilité est une tâche vraisemblablement complexe et un ajout de bruit sélectionné dans le pire des cas risque de détruire l'utilité des données générées.

Alternativement, Abadi *et al.* (2016) proposent de bruiter le gradient durant l'entraînement des réseaux de neurones par l'algorithme de descente du gradient stochastique (SGD). L'adaption qu'ils proposent de SDG est illustrée par l'algorithme 3. À chacune de ses étapes, l'algorithme calcule le gradient pour un sous-ensemble aléatoire d'exemples, coupe la norme de chaque gradient, calcule la moyenne des gradients puis y ajoute du bruit avant de prendre un pas dans la direction opposée de ce gradient bruité. Majorer la norme du gradient sert à limiter l'influence de chaque exemple d'entraînement sur l'apprentissage pour contrôler la sensibilité tandis que le bruit additif préserve la confidentialité. En fonction de la sensibilité choisie et du bruit ajouté, des propriétés de composition avancées de la confidentialité différentielle (Dwork et Roth, 2014; Abadi *et al.*, 2016; Mironov, 2017) majorent le budget de confidentialité (ϵ, δ) accumulé.

Cette version de SGD rend notamment possible l'entraînement de GANs de façon à ce que leurs poids satisfassent la DP (Xu *et al.*, 2019a). Ainsi, des données synthétiques qui respectent la DP peuvent être générées.

Algorithme 3 Adaptation de SGD différentiellement confidentielle. Les entrées de l'aglorithme sont les paramètres du réseau θ; la fonction de coût $\mathcal{L}$; le nombre d'itérations T; le taux d'apprentissage η; l'échelle de bruit σ; la taille d'échantillonnage L; la borne sur la norme du gradient C.

DP-SGD($\{x_1, \ldots, x_N\}$, θ, $\mathcal{L}(\theta, x_i)$, T, η, σ, L, C)
  for $t \in [1, \ldots, T]$ **do**
    $L_t \leftarrow$ ÉCHANTILLONNER($\{x_1, \ldots, x_N\}$, L)
    for $i \in L_t$ **do**
      $g_t(x_i) \leftarrow \nabla_{\theta_t} \mathcal{L}(\theta_t, x_i)$
      $\bar{g}_t(x_i) \leftarrow g_t(x_i) / \max(1, \frac{\|g_t(x_i)\|_2}{C})$
    end for
    $\tilde{g}_t \leftarrow \frac{1}{L} \sum_i \bar{g}_t(x_i) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 C^2 I)$
    $\theta_{t+1} \leftarrow \theta_t - \eta \tilde{g}_t$
  end for
  $(\epsilon, \delta) \leftarrow$ BUDGET( )
return θ_T, (ϵ, δ)
end

Chapitre 5

Désensibilisation de Adult Dataset

L'objectif principal du stage de recherche est de développer une méthode de désensibilisation et de s'assurer que les données qu'elle génère conservent une utilité comparable à celle des données sensibles. Lorsque la publication de données qui respecte la confidentialité remplace des données sensibles où elles servent pour une tâche déterminée, l'utilité des données générées peut être évaluée d'après cette tâche. Cette section présente le scénario fictif dans lequel le jeu de données publiques Adulte (Dua et Graff, 1994), présumé sensible, doit être désensibilisé pour développer un classificateur qui prédit si le revenu personnel est au-dessus ou au-dessous de 50 000$ par an.

Adulte est un jeu de 32 561 données tabulaire provenant du recensement américain de 1994 ayant 15 attributs dont 5 sont numériques. La table 5.1 décrit les variables telles qu'elles apparaissent dans le jeu de données.

Attribut	**Type**	**Catégories**
age	numérique	$(16.927, 24.3]$, $(24.3, 31.6]$, $(31.6, 38.9]$, $(38.9, 46.2]$, $(46.2, 53.5]$, (53.5,60.8] (60.8,68.1] $(68.1, 75.4]$ (75.4,82.7] (82.7,90]
workclass	catégorielle	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
fnlwgt	numérique	$(10812.58, 159527.0]$, $(159527.0, 306769.0]$, $(306769.0, 454011.0]$, $(454011.0, 601253.0]$, $(601253.0, 748495.0]$, $(748495.0, 895737.0]$, $(895737.0, 1042979.0]$, $(1042979.0, 1190221.0]$, $(1190221.0, 1337463.0]$, $(1337463.0, 1484705.0]$
education	catégorielle	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool

Attribut	Type	Catégories
education-num	catégorielle	13, 9, 7, 14, 5, 10, 12, 11, 4, 16, 15, 3, 6, 2, 1, 8
marital-status	catégorielle	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
occupation	catégorielle	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces
relationship	catégorielle	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
race	catégorielle	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
sex	catégorielle	Female, Male
capital-gain	numérique	$(-99.999, 9999.9]$, $(9999.9, 19999.8]$, $(19999.8, 29999.7]$, $(29999.7, 39999.6]$, $(39999.6, 49999.5]$, $(49999.5, 59999.4]$, $(59999.4, 69999.3]$, $(69999.3, 79999.2]$, $(79999.2, 89999.1]$, $(89999.1, 99999.0]$
capital-loss	numérique	$(-4.356, 435.6]$, $(435.6, 871.2]$, $(871.2, 1306.8]$, $(1306.8, 1742.4]$, $(1742.4, 2178.0]$, $(2178.0, 2613.6]$, $(2613.6, 3049.2]$, $(3049.2, 3484.8]$, $(3484.8, 3920.4]$, $(3920.4, 4356.0]$
hours-per-week	numérique	$(0.902, 10.8]$, $(10.8, 20.6]$, $(20.6, 30.4]$, $(30.4, 40.2]$, $(40.2, 50.0]$, $(50.0, 59.8]$, $(59.8, 69.6]$, $(69.6, 79.4]$, $(79.4, 89.2]$, $(89.2, 99.0]$
native-country	catégorielle	United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy,Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands
income	catégorielle	$> 50k$, $\leq 50k$

Tableau 5.1. Description des attributs du jeu de données Adulte. Les variables numériques ont été discrétisées chacune en 10 catégories.

Cette section compare deux approches de désensibilisation de données, soit la génération de données synthétiques et le k-anonymat. Plus précisément, les données générées par CTGAN sont comparées à celles produites par l'algorithme Mondrian. La structure et les hyperparamètres de CTGAN sont ceux suggérés par défaut par Xu *et al.* (2019b). Pour Mondrian, comme un nombre exponentiel de combinaisons de dimensions peut être utilisé pour effectuer des attaques par liaison d'attributs, toutes les variables sont considérées comme étant quasi-identifiantes. De plus, tel que l'algorithme est décrit à la section 4.1, à chacune de ses itérations, la dimension choisie pour séparer les données est celle ayant le plus grand diamètre normalisé et la valeur de séparation pour les variables numériques est la médiane. Une fois les données partitionnées, les classes d'équivalences sont généralisées: au sein d'un groupe, les variables numériques sont remplacées par la moyenne et les variables catégorielles par le mode.

Pour évaluer la qualité des données générées, nous souhaitons savoir si elles permettent de prédire si le revenu annuel est au-dessus ou au-dessous de 50 000$. Pour ce faire, les données sensibles sont premièrement séparées en ensembles de test et d'entraînement. Ensuite, les données d'entraînement sont désensibilisées par CTGAN et Mondrian pour plusieurs valeurs de k. Pour chaque version désensibilisée des données d'entraînement, un modèle est entraîné à prédire le revenu. Leurs performances à prédire le revenu sur l'ensemble test initialement isolé sont comparées à celle d'un modèle de référence entraîné sur l'ensemble d'entraînement sensible. Plus la performance d'un prédicteur entraîné sur des données désensibilisées est proche de celle du modèle de référence, plus nous supposons que l'utilité de ces données est préservée. Les performances sont également comparées à celle d'un jeu de données généré en échantillonnant indépendamment chaque dimension à partir de l'ensemble d'entraînement brisant ainsi les relations inter-dimensionnelles. Cet ensemble de données aux marginales indépendantes est considéré comme une référence minorant l'utilité. Évaluer la performance sur un ensemble de test sensible permet d'estimer si les modèles entraînés sur des données désensibilisées devraient performer dans un environnement de production où les données sont sensibles.

Les classificateurs implémentés pour prédire le revenu sont des arbres de décision (Loh, 2011). La métrique de performance est le score $f1$ pondéré par le support des classes dans l'ensemble de test. Le score $f1$ est défini comme la moyenne harmonique de la précision et du rappel et quantifie le compromis entre ceux-ci:

$$f1 = 2 \times \frac{\text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}}$$

où

$$\text{précision} = \frac{VP}{VP + FP} \text{ et } \text{rappel} = \frac{VP}{VP + FN}$$

avec VP le nombre de vrais positifs, FP de faux positifs et FN de faux négatifs. La précision est une mesure d'exactitude qui correspond à la proportion des items pertinents parmi l'ensemble des items proposés. Le rappel (ou sensibilité) est une mesure d'exhaustivité. Le rappel est la proportion des items pertinents proposés parmi l'ensemble des items pertinents. C'est le taux de bonnes classifications par classe.

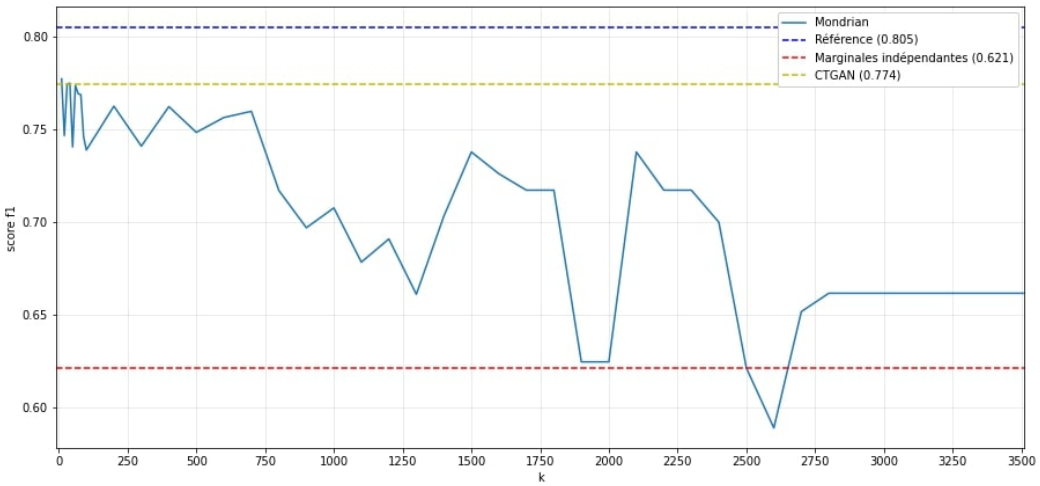


Fig. 5.1. Score $f1$ pour la prédiction du revenu annuel pour CTGAN et l'algorithme de Mondrian en fonction de différentes valeurs de k

Dans l'éventualité où les données sensibles ne servent pas nécessairement à une tâche prédéterminée comme prédire le revenu, ce score se généralise par ce qu'on appelle la prédiction dimensionnelle (DWP, de l'anglais *Dimension-wise Prediction*) (Choi *et al.*, 2017; Xu *et al.*, 2019b). Il s'agit de prédire chaque dimension successivement et de calculer la moyenne des scores $f1$ ainsi obtenus. Pour obtenir un score $f1$ sur les variables numériques, celles-ci ont été discrétisées chacune en 10 catégories avant d'entraîner les modèles, mais après la désensibilisation. Ces catégories sont énumérées dans la dernière colonne de la table 5.1. Le score d'utilité DWP rapporté est 1 moins la différence entre les moyennes des scores $f1$ du modèle de référence et du modèle entraîné sur les données désensibilisées. Ce score d'utilité est positif et un score de 1 représente la performance de la référence.

La figure 5.1 graphe les scores $f1$ des classificateurs entraînés à prédire le revenu sur les jeux de données désensibilisés pour différentes valeurs de k avec l'algorithme Mondrian (courbe bleue pâle). Similairement, la figure 5.2 illustre les scores d'utilité DWP en fonction de k. Sur chacun des graphes, la courbe horizontale bleue marque la performance du classificateur entraîné sur l'ensemble d'entraînement sensible qui représente une borne supérieure d'utilité. Inversement, la courbe horizontale rouge marque une borne inférieure d'utilité

générée par un classificateur entraîné sur les données aux marginales indépendantes. Finalement, la courbe horizontale jaune marque la performance du classificateur entraîné sur les données générées par CTGAN.

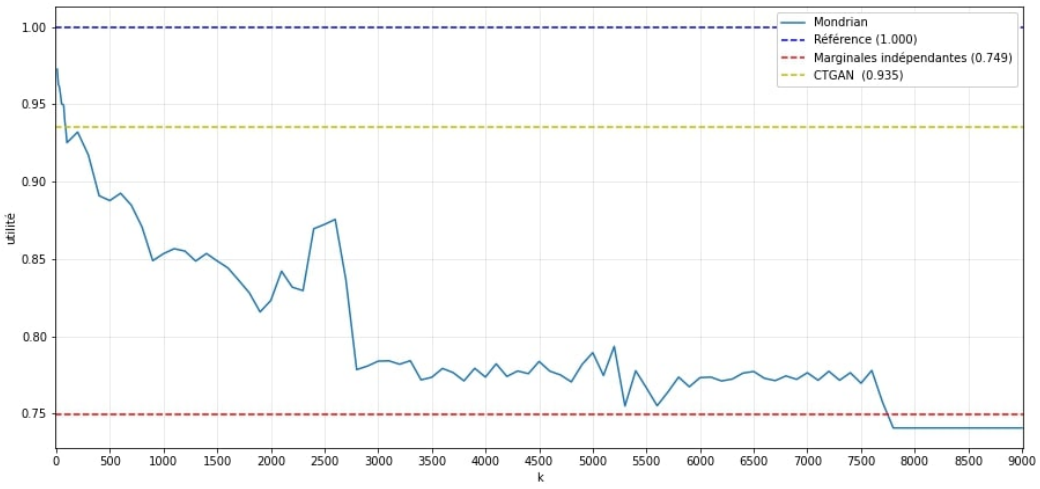


Fig. 5.2. Score d'utilité DWP pour CTGAN et l'algorithme de Mondrian en fonction de différentes valeurs de k

Plus k est élevé, plus les données sont anonymisées. Les graphiques 5.1 et 5.2 montrent que l'utilité aux sens de la prédiction du revenu et de DWP décroît effectivement en fonction de k pour l'algorithme de Mondrian. C'est l'expression du compromis entre l'utilité et la confidentialité: un jeu de données ne peut être complètement désensibilisé et demeurer utile. Le choix de k revient à ceux qui veulent publier des données en fonction de leurs objectifs d'utilité et de confidentialité. Par exemple, pour $k = 100$, le risque de ré-identification est majoré par 1% et le score $f1$ est de 0.739 et le score d'utilité DWP est de 0.925. Toutefois, de grandes valeurs de k donnent lieu à des pertes inacceptables d'utilité. Pour $k = 2000$, l'algorithme de Mondrian partionne Adulte en seulement 9 classes d'équivalence dont la taille moyenne est de 2894 enregistrements.

CTGAN conserve une utilité comparable aux meilleures performances de Mondrian. Bien que, contrairement au k-anonymat, CTGAN ne fourni aucune garantie formelle de confidentialité, CTGAN produit des données entièrement synthétiques si bien qu'il n'existe pas de bijection entre les données générées et les données sensibles, ce qui protège contre les attaques par liaison d'attributs et limite le risque de ré-identification.

Chapitre 6

Conclusion

Deux directions de recherches dont l'intersection est la génération de données ont été empruntées par ce mémoire. La première, académique, place la question de la synthèse de population dans un cadre général. Est-il possible de transférer la modélisation des dépendences multi-variées d'une population source à une population cible dont seules les informations marginales sont connues en autant que ces populations partagent des caractéristiques similaires de dépendences? Lorsque la population source est un sous-ensemble de la population cible ou que ces populations proviennent de régions distinctes mais similaires, nous avons vu que les copules permettent effectivement de générer des données pour la cible en utilisant ses distributions marginales et un échantillon de la population source. Cependant, en pratique, identifier une copule appropriée est une tâche potentiellement complexe, et ce, particulièrement dans le cas de distributions multivariées discrètes pour lesquelles la copule n'est pas nécessairement unique. Dans ces circonstances, le choix d'une copule peut affecter l'exactitude du transfert des dépendences d'une distribution à l'autre car la copule de l'une n'est peut-être pas partagée par l'autre. Jusqu'à quelle mesure l'hypothèse selon laquelle deux populations partagent une même copule tient et selon quelles conditions peut-elle être émise? Voilà un questionnement qui mérite une réflexion formelle plus approfondie qu'elle ne l'a été dans ce mémoire.

La seconde direction, entreprise dans le cadre d'un stage Mitacs en industrie, est celle de la publication de données qui préserve la confidentialité. À défaut de révéler la méthodologie relative aux besoins spécifiques d'affaire, on a discuté différents cadres de confidentialité: le k-anonymat, les données synthétiques et la confidentialité différentielle, et présenté une analyse rudimentaire de la qualité des données dans le contexte de désensibilisation du jeu de données publique Adulte. Les données générées par CTGAN et l'algorithme Mondrian conservent une utilité relativement aux données sensibles lorsqu'elles sont utilisées pour développer des modèles prédictifs. Toutefois, ces méthodes sont fondées sur des définitions de confidentialité différentes et difficilement comparables, ce qui rend l'analyse du compromis entre l'utilité

et la confidentialité difficile. Pour une utilité fixe, quelle méthode est la plus confidentielle? Des mesures de confidentialité *a posteriori*, agnostiques du cadre de confidentialité et qui ciblent différents risques, comme les risques de ré-identification et de divulgation d'attributs pourraient aider à répondre à cette question et comparer les différentes méthodes populaires de la littérature de publication de données qui préserve la confidentialité.

Références bibliographiques

Martin Abadi, Andy Chu, Ian Goodfellow, Brendan McMahan, Ilya Mironov, Kunal Talwar et Li Zhang : Deep learning with differential privacy. *In 23rd ACM Conference on Computer and Communications Security (ACM CCS)*, pages 308–318, 2016.

Charu C. Aggarwal : On k-anonymity and the curse of dimensionality. *In Proceedings of the 31st International Conference on Very Large Data Bases*, VLDB '05, pages 901–909. VLDB Endowment, 2005.

Theo Arentze, Harry J. P. Timmermans et Frank Hofman : Creating synthetic household populations: Problems and approach. *Transportation Research Record*, 2014:85–91, 2007.

Theo A. Arentze et Harry J.P. Timmermans : A learning based transportation oriented simulation system. *Transportation Research Part B*, 38(7):613–633, 2004.

Joshua Auld, Mahmoud Javanmardi et Abolfazl Mohammadian : Integration of activity scheduling and traffic assignment in ADAPTS activity-based model. *In TRB 91st Annual Meeting Compendium of Papers DVD*, numéro 12-4225, Washington DC, United States, janvier 2012. Transportation Research Board.

Joshua Auld, Abolfazl Mohammadian et Kermit Wies : Population synthesis with subregion-level control variable aggregation. *Journal of Transportation Engineering*, 135 (9):632–639, 2009.

Joshua Auld et Kermit Wies : Population synthesis with subregion-level control variable aggregation. *Journal of Transportation Engineering-asce*, 135, 2009.

Athanassios Avramidis, Nabil Channouf et Pierre L'Ecuyer : Efficient correlation matching for fitting discrete multivariate distributions with arbitrary marginals and normal-copula dependence. *INFORMS Journal on Computing*, 21:88–106, 2009.

Johan Barthelemy et Philippe L. Toint : Synthetic population generation without a sample. *Transportation Science*, 47(2):266–279, 2013.

Richard J. Beckman, Keith A. Baggerly et Michael D. McKay : Creating synthetic baseline populations. *Transportation Research Part A*, 30(6):415–429, 1996.

Steven Bellovin, Preetam Dutta et Nathan Reitinger : Privacy and synthetic datasets. *Stanford Technology Law Review*, 22, 09 2018.

Jon Louis Bentley : Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, 1975.

Chandra R. Bhat et Naveen Eluru : A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B*, 43 (7):749–765, 2009.

Christopher M. Bishop : *Pattern Recognition and Machine Learning*. Springer, New-York, NY, USA, 2006.

Guillem Boquet, Antoni Morell, Javier Serrano et Jose Lopez Vicario : A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection. *Transportation Research Part C*, 115:102622, 2020.

Edoardo Borgomeo, Georg Pflug, Jim W. Hall et Stefan Hochrainer-Stigler : Assessing water resource system vulnerability to unprecedented hydrological drought using copulas to characterize drought duration and deficit. *Water Resources Research*, 51(11): 8927–8948, 2015.

Kathryn Born, Shamsunnahar Yasmin, Daehyun You, Naveen Eluru, Chandra R. Bhat et Ram M. Pendyala : Joint model of weekend discretionary activity participation and episode duration. *Transportation Research Record*, 2413(1):34–44, 2014.

Stanislav S. Borysov et Jeppe Rich : Introducing synthetic pseudo panels: application to transport behaviour dynamics. *Transportation*, 48(5):2493–2520, 2021.

Stanislav S. Borysov, Jeppe Rich et Francisco C. Pereira : How to generate micro-agents? A deep generative modeling approach to population synthesis. *Transportation Research Part C*, 106:73–97, 2019.

John L. Bowman, Mark Bradley, Joe Castiglione et Supin L. Yoder : Making advanced travel forecasting models affordable through model transferability. *In Transportation Research Board 93rd Annual Meeting*, janvier 2014.

Mark Bradley, John L. Bowman et Bruce Griesenbeck : Sacsim: An applied activity-based model system with fine-level spatial and temporal resolution. *Journal of Choice Modelling*, 3(1):5–31, 2010.

Daniele Casati, Kirill Müller, Pieter J Fourie, Alexander Erath et Kay W Axhausen : Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record*, 2493(1):107–116, 2015.

Umberto Cherubini, Elisa Luciano et Walter Vecchiato : *Copula Methods in Finance*. John Wiley & Sons, Chichester, United Kingdom, 2004.

Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F. Stewart et Jimeng Sun : Generating multi-label discrete patient records using generative adversarial networks. *In* Finale Doshi-Velez, Jim Fackler, David Kale, Rajesh Ranganath,

Byron Wallace et Jenna Wiens, éditeurs : *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 de *Proceedings of Machine Learning Research*, pages 286–305. PMLR, Aug 2017.

Abdoul-Ahad Choupani et Amir Reza Mamdoohi : Population synthesis using iterative proportional fitting (IPF): A review and future research. *Transportation Research Procedia*, 17:223–233, 2016. International Conference on Transportation Planning and Implementation Methodologies for Developing Countries (12th TPMDC) Selected Proceedings, IIT Bombay, Mumbai, India, 10-12 December 2014.

Gregory F. Cooper : The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2):393–405, 1990.

Yves-Alexandre de Montjoye, Laura Radaelli, Vivek Kumar Singh et Alex Sandy Pentland : Identity and privacy. unique in the shopping mall: on the reidentifiability of credit card metadata. *Science (New York, N.Y.)*, 347(6221):536–539, 2015.

W. Edwards Deming et Frederick F. Stephan : On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

Dheeru Dua et Casey Graff : UCI machine learning repository, 1994.

Gérald Duguay, Woo Jung et Daniel L. McFadden : SYNSAM: A methodology for synthesizing household transportation survey data. Working paper 7618, Institute of Transportation Studies, University of California, Berkeley, CA, USA, 1976.

Cynthia Dwork : A firm foundation for private data analysis. *Communications of the ACM*, 2011.

Cynthia Dwork, Frank McSherry, Kobbi Nissim et Adam Smith : Calibrating noise to sensitivity in private data analysis. *In* Shai Halevi et Tal Rabin, éditeurs : *Theory of Cryptography*, pages 265–284, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

Cynthia Dwork et Aaron Roth : The algorithmic foundations of differential privacy. *Foundations and trends in theoretical computer science*, 9(3–4):211–407, 2014.

Khaled El Emam, Fida Kamal Dankar, Romeo Issa, Elizabeth Jonker, Daniel Amyot, Elise Cogo, Jean-Pierre Corriveau, Mark Walker, Sadrul Chowdhury, Regis Vaillancourt, Tyson Roffey et Jim Bottomley : A globally optimal k-anonymity method for the de-identification of health data. *Journal of the American Medical Informatics Association*, 16(5):670–682, 2009.

Naveen Eluru, Abdul Rawoof Pinjari, Jessica Y. Guo, Ipek Nese Sener, Sivaramakrishnan Srinivasan, Rachel B. Copperman et Chandra R. Bhat : Population updating system structures and models embedded in the comprehensive econometric microsimulator for urban systems. *Transportation Research Record*, 2076:171–182, 2008.

Bilal Farooq, Michel Bierlaire, Ricardo Hurtubia et Gunnar Flötteröd : Simulation based population synthesis. *Transportation Research Part B*, 58:243–263, 2013.

Sergio Garrido, Stanislav S. Borysov, Francisco C. Pereira et Jeppe Rich : Prediction of rare feature combinations in population synthesis: Application of deep generative modelling. *Transportation Research Part C*, 120:102787, 2020.

Christian Genest et Anne-Catherine Favre : Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of hydrologic engineering*, 12(4):347–368, 2007.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville et Yoshua Bengio : Generative adversarial nets. *In* Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence et K.Q. Weinberger, éditeurs : *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Lukas Günthermann, Ivor Simpson et Daniel Roggen : Smartphone location identification and transport mode recognition using an ensemble of generative adversarial networks. *In Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp-ISWC '20, pages 311—316, New York, NY, USA, septembre 2020. Association for Computing Machinery.

Jessica Guo et Chandra Bhat : Population synthesis for microsimulating travel behavior. *Transportation Research Record*, 2014:92–101, 2007.

David Heckerman, Dan Geiger et David M. Chickering : Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20(3):194–243, 1995.

Zengyi Huang et Paul Williamson : A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata. Working Paper 2001/2, University of Liverpool, octobre 2001.

Sebastian Hörl et Milos Balac : Synthetic population and travel demand for Paris and Île-de-France based on open and publicly available data. *Transportation Research Part C*, 130:103291, 2021.

Vijay S. Iyengar : Transforming data to satisfy privacy constraints. *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '02, pages 279–288, New York, NY, USA, 2002. Association for Computing Machinery.

Amel Jaoua, Pierre L'Ecuyer et Louis Delorme : Call-type dependence in multiskill call centers. *SIMULATION*, 89(6):722–734, 2013.

Byungduk Jeong, Wonjoon Lee, Deok-Soo Kim et Hayong Shin : Copula-based approach to synthetic population generation. *PLOS ONE*, 11:1–28, 2016.

Harry Joe : *Multivariate Models and Dependence Concepts*. Springer, New York, NY, USA, 1997.

Harry Joe : *Dependence modeling with copulas*. CRC Press, Boca Raton, FL, USA, 2015.

Shih-Chieh Kao, Hoe Kyoung Kim, Cheng Liu, Xiaohui Cui et Budhendra L. Bhaduri : Dependence-preserving approach to synthesizing household characteristics. *Transportation Research Record*, 2302:192–200, 2012.

Kartik Kaushik, Cinzia Cirillo et Fabian Bastin : On modelling human population characteristics with copulas. *Procedia Computer Science*, 151:210–217, 2019. The 10th International Conference on Ambient Systems, Networks and Technologies (ANT 2019) / The 2nd International Conference on Emerging Data and Industry 4.0 (EDI40 2019) / Affiliated Workshops.

Diederik P. Kingma et Max Welling : Auto-encoding variational bayes. *In International Conference on Learning Representations (ICLR)*, Banff, AB, Canada, avril 2014.

Karthik Konduri, Daehyun You, Venu Garikapati et Ram Pendyala : Enhanced synthetic population generator that accommodates control variables at multiple geographic resolutions. *Transportation Research Record*, 2563:40–50, 2016.

Wai Lam et Fahiem Bacchus : Learning bayesian belief networks: An approach based on the mdl principle. *Computational Intelligence*, 10(3):269–293, 1994.

Daniel T. Larose et Chantal D. Larose : *Data Preprocessing*, chapitre 2, pages 16–50. John Wiley & Sons, Hoboken, NJ, USA, 2014.

Kristen LeFevre, David J. DeWitt et Raghu Ramakrishnan : Mondrian multidimensional k-anonymity. *In 22nd International Conference on Data Engineering (ICDE'06)*, pages 25–25, 2006.

Jun-Lin Lin et Meng-Cheng Wei : An efficient clustering method for k-anonymization. *In Proceedings of the 2008 International Workshop on Privacy and Anonymity in Information Society*, PAIS '08, pages 46–50, New York, NY, USA, 2008. Association for Computing Machinery.

Wei-Yin Loh : Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1:14 – 23, 2011.

Adam Meyerson et Ryan Williams : On the complexity of optimal k-anonymity. *In Proceedings of the Twenty-Third ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '04, pages 223–228, New York, NY, USA, 2004. Association for Computing Machinery.

Ilya Mironov : Rényi differential privacy. *In 2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pages 263–275, 2017.

Kirill Müller et Kay W. Axhausen : Population synthesis for microsimulation: State of the art. *In TRB 90th Annual Meeting Compendium of Papers DVD*, numéro 11-1789, Washington DC, United States, janvier 2011. Transportation Research Board.

Arvind Narayanan et Vitaly Shmatikov : Robust de-anonymization of large sparse datasets. *In 2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125,

2008.

Roger B Nelsen : *An Introduction to Copulas*. Springer, New York, NY, USA, second édition, 2006.

Ostap Okhrin, Alexander Ristig et Ya-Fei Xu : *Copulae in High Dimensions: An Introduction*, chapitre 13, pages 247–277. Springer, Berlin, Germany, third édition, 2017.

Boris Oreshkin, Nazim Réegnard et Pierre L'Ecuyer : Rate-based daily arrival process models with application to call centers. *Operations Research*, 64, 2016.

Evangelos Panos et Stavroula Margelou : Long-term solar photovoltaics penetration in single- and two-family houses in switzerland. *Energies*, 12(13), 2019.

Abdul Rawoof Pinjari, Chandra R. Bhat et David A. Hensher : Residential self-selection effects in an activity time-use behavior model. *Transportation Research Part B*, 43(7):729–748, 2009.

David R. Pritchard et Eric J. Miller : Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, 39(3):685–704, 2012.

Tejsingh A. Rana, Sujan Sikder et Abdul Rawoof Pinjari : Copula-based method for addressing endogeneity in models of severity of traffic crash injuries: Application to two-vehicle crashes. *Transportation Research Record*, 2147:75–87, 2010.

Ismaïl Saadi, Ahmed Mustafa, Jacques Teller et Mario Cools : Forecasting travel behavior using markov chains-based approaches. *Transportation Research Part C*, 69:402–417, 2016a.

Ismaïl Saadi, Ahmed Mustafa, Jacques Teller, Bilal Farooq et Mario Cools : Hidden markov model-based population synthesis. *Transportation Research Part B*, 90:1–21, 2016b.

Paul Salvini et Eric J. Miller : ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems. *Networks and spatial economics*, 5(2):217–234, 2005.

Sujan Sikder, Abdul Rawoof Pinjari, Sivaramakrishnan Srinivasan et Nowrouzian Roosbeh : Spatial transferability of travel forecasting models: a review and synthesis. *International Journal of Advances in Engineering Sciences and Applied Mathematics*, 5(2):104–128, 2013.

Abe Sklar : Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8:229–231, 1959.

Laron Smith, Richard Beckman, Keith Baggerly, Doug Anson et Michael Williams : TRANSIMS: Transportation analysis and simulation system. techreport LA-UR-95-1641, Los Alamos National Laboratory, Los Alamos, NM, USA, juillet 1995.

Lijun Sun et Alexander Erath : A Bayesian network approach for population synthesis. *Transportation Research Part C*, 61:49–62, 2015.

Lijun Sun, Alexander Erath et Ming Cai : A hierarchical mixture modeling framework for population synthesis. *Transportation Research Part B*, 114:199–212, 2018.

Latanya Sweeney : k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

Paul Waddell : UrbanSim: Modeling urban development for land use, transportation, and environmental planning. *Journal of the American Planning Association*, 68(3):297–314, 2002.

Chugui Xu, Ju Ren, Deyu Zhang, Yaoxue Zhang, Zhan Qin et Kui Ren : Ganobfuscator: Mitigating information leakage under gan via differential privacy. *IEEE Transactions on Information Forensics and Security*, 14(9):2358–2371, 2019a.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante et Kalyan Veeramachaneni : Modeling tabular data using conditional GAN. *In* Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc et Emily B. Fox, éditeurs : *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 7335–7345, Red Hook, NY, USA, 2019b. Curran Associates Inc.

Ali Yazdizadeh, Zachary Patterson et Bilal Farooq : Semi-supervised GANs to infer travel modes in GPS trajectories. *Journal of Big Data Analytics in Transportation*, 3 (3):201–211, 2021.

Xin Ye, Karthik Konduri, Ram Pendyala, Bhargava Sana et Paul Waddell : Methodology to match distributions of both household and person attributes in generation of synthetic populations. *In TRB 88th Annual Meeting Compendium of Papers DVD*, numéro 09-2096, Washington DC, United States, jan 2009. Transportation Research Board.

Mogeng Yin, Madeleine Sheehan, Sidney Feygin, Jean-François Paiement et Alexei Pozdnoukhov : A generative model of urban activities from cellular data. *IEEE Transactions on Intelligent Transportation Systems*, 19(6):1682–1696, 2018.

Danqing Zhang, Junyu Cao, Sid Feygin, Dounan Tang, Zuo-Jun(Max) Shen et Alexei Pozdnoukhov : Connected population synthesis for transportation simulation. *Transportation Research Part C*, 103:1–16, 2019.

Dominik Ziemke, Johan W. Joubert et Kai Nagel : Accessibility in a post-apartheid city: Comparison of two approaches for accessibility computations. *Networks and Spatial Economics*, 18(2):241–271, 2018.