Analyse comparée de méthodes de générations de données hétérogènes, multidimensionnelles et réalistes

Michael Corbeau*, Emmanuelle Claeys*, Mathieu Serrurier*, Pascale Zaraté*

*Institut de recherche en informatique de Toulouse
Cr Rose Dieng-Kuntz, 31400 Toulouse, France
Prénom.Nom@irit.fr

1 Introduction

La planification stratégique et l'optimisation des ressources dans des domaines critiques nécessitent des données synthétiques réalistes. Cependant, générer des données capables de capturer les complexités des données réelles, tout en reflétant leurs déséquilibres, représente un défi considérable en raison des hétérogénéités et des corrélations spécifiques des données réelles. L'étude compare plusieurs techniques de génération, notamment les GANs, les Variational Autoencoders et les modèles de diffusion, pour évaluer leur capacité à reproduire la diversité et la complexité des données tout en respectant les contraintes de variabilité.

2 Contexte et Défis

Nos données sont issues de données réelles d'interventions de pompiers qui ne suivent pas de distributions statistiques classiques. Les variables incluent des coordonnées géographiques, des aspects temporels (heure, jour, mois), des durées d'intervention et des types d'incidents.

L'objectif est de créer un simulateur d'interventions capable de produire des données synthétiques qui respectent la variabilité et les caractéristiques des données réelles. Cela implique de préserver les anomalies statistiques, les biais observés et les corrélations fines tout en générant des volumes suffisants pour répondre aux besoins de modélisation.

3 Méthodes évaluées

Plusieurs approches sont examinées pour évaluer leur capacité à générer des données tabulaires réalistes :

Méthodes de base : des techniques simples comme l'échantillonnage aléatoire ou le mélange avec remplacement servent de points de comparaison.

Tabular Variational Autoencoders (TVAE) : ce modèle encode les données dans un espace latent avant de les reconstruire. Bien qu'efficace pour certaines tâches, le TVAE montre des limites dans la préservation des corrélations complexes.

Generative Adversarial Networks (GANs) : les GANs, composés d'un générateur et d'un discriminateur, se distinguent dans la génération d'images. Toutefois, leur performance diminue face aux données tabulaires en raison des structures de corrélation différentes.

Modèle de diffusion : ces modèles ajoutent du bruit progressif aux données pour apprendre à les reconstruire. Ils se révèlent particulièrement adaptés aux données mixtes et tabulaires, capturant mieux la diversité et les caractéristiques complexes.

4 Critères d'évaluation

Pour comparer ces méthodes, des métriques standard et spécifiques au domaine sont utilisées :

— **Distance de Wasserstein et Maximum Mean Discrepancy (MMD)** : Ces métriques évaluent la similarité entre les distributions réelles et synthétiques.
— **Density et Coverage :** Elles mesurent la représentation locale et la diversité des données synthétiques.
— **Métriques spécifiques au domaine :** Ces mesures incluent la distribution spatio-temporelle, la variabilité du nombre d'interventions par secteur et les corrélations fines entre incidents et mois.

5 Résultats principaux

Les résultats mettent en évidence des différences significatives entre les approches :

— Les méthodes de base es techniques échouent à capturer les complexités et corrélations observées dans les données réelles.
— Le VAE biaise souvent les données en se concentrant sur les modes les plus fréquents.
— Les GANs montrent une certaine efficacité dans la préservation des distributions globales mais peinent à respecter les contraintes spécifiques du domaine, notamment les corrélations incident/mois et la variabilité du nombre d'interventions par secteur
— Les modèles de Diffusion (TabDiff et TinyDiff) surpassent les autres selon toutes les métriques évaluées. TabDiff, en particulier, excelle dans la préservation des corrélations fines et des anomalies statistiques. Il génère des données synthétiques qui respectent les biais géographiques et les distributions temporelles, tout en offrant une variabilité réaliste.

6 Conclusion

Les modèles de diffusion, en particulier TabDiff et TinyDiff, sont les plus performants pour générer des données synthétiques réalistes à partir d'un jeu de données de petite taille. Ces modèles préservent les corrélations complexes, les anomalies statistiques et la variabilité des données réelles, tout en respectant les besoins de scalabilité. TabDiff, grâce à l'utilisation d'une variables cible, reproduit fidèlement les tendances et corrélations critiques pour des simulations fiables. Ces résultats valident leur application pour des scénarios réels et ouvrent des perspectives dans d'autres domaines nécessitant des données synthétiques robustes.