Generating Synthetic Geolocation Data using Conditional Tabular Generative Adversarial Networks: Issues and Challenges

Abu Yousuf Md Abdullah[1] **Bing Hu**[2] **Mohammad Ahmed Basri**[3] **Shu-Feng Tsao**[4] **Zahid Butt**[4] **Helen Chen**[2,4,5]
[1]School of Planning, [2]Cheriton School of Computer Science, [3]Department of Systems Design Engineering
[4]School of Public Health Sciences, [5]Department of Statistics and Actuarial Science
University of Waterloo
{aymabdullah, bingxu.hu, mabasri, s7tsao, zahid.butt, helen.chen}@uwaterloo.ca

Abstract

The advent of generative adversarial neural networks (GANs) has created a new frontier for generating synthetic datasets having the same statistical properties as the real data from which it was generated. Real data comprises various types of data, which further complicates the synthesis process as specialized GANs are required to handle these data with diverse variable types. In this context, Conditional Tabular GAN (CTGAN) can be very useful as it has been found to handle various types of non-spatial data successfully. However, the use of CTGAN in generating spatial or geolocation data has remained largely unexplored. This study uses the Traffic Collisions Open Data from the Toronto Police Service to demonstrate the challenges involved in modeling geolocation data in CTGAN and reports the potential limitations of the deep learning data synthesizer in generating synthetic datasets with substantial geolocation and spatial components.

1 Introduction

Deep learning data synthesizers, such as Generative Adversarial Networks (GANs), allow privacy-sensitive synthetic datasets to be developed [1–3]. As a majority of the datasets collected in real life are tabular in nature, the use of Conditional Tabular GANs (CTGANs) allows considerable flexibility in exploiting these datasets. The application of CTGAN offers several advantages, which include (but are not limited to) (i) handling both continuous and discrete data, (ii) accounting for imbalances, and (iii) multi-modality in the data [2]. Although geolocation is an important component of most real-life datasets, studies exploring the suitability of existing GANs in generating geolocation synthetic data are limited. Furthermore, due to the high privacy risk concerns associated with geolocation data, using GANs can offer a substantial advantage over conventional statistical and machine learning techniques that generate synthetic data from simply modeling the distribution of the real dataset or adopt decision-based approaches [1, 4]. For this purpose, the introduction of random noise in the generator and the validation through discriminator during the data synthesis process can offer added privacy protections, which otherwise, are not available in conventional models [2, 4].

As a precautionary measure, while publicly sharing datasets, the geolocation information in deidentified and anonymized datasets can often be deliberately distorted to protect privacy. For example, while sharing the events data (such as crimes or collisions), Toronto Police deliberately offsets the occurrences to the nearest road intersection node to protect the privacy of parties involved in the occurrence [5]. This could be severely problematic for advanced analyses and research as risk modeling in various use cases involves modeling the events of interest with high precision and granularity to accurately identify the putative the risk factors and adjust for any unmeasured or latent confounders [6, 7].

Unfortunately, modeling geolocation information, such as latitude and longitude values, in CTGAN could be challenging as these two variables must be considered conjointly during the synthesis process. For example, a set of latitude and longitude values should be interpreted together to define a single point location. Moreover, this joint modeling must be done while considering their (spatial) relationship with other variables in the dataset. Furthermore, as most GANs were developed with a focus on analyzing image and tabular datasets [2, 8, 9], there are limitations in modeling geolocation data, which often require specific geographic and projected coordinate systems to define and characterize them in deep learning processes properly.

Therefore, considering existing research gaps, this study aims to utilize the Traffic Collisions Open Data from the Toronto Police Service to demonstrate the issues involved in modeling geolocation data in CTGAN and discusses the potential limitations of the deep learning data synthesizer that should be considered in generating geolocation data.

2 Methods

2.1 Study Area and Data

The study considers the City of Toronto in Canada (Fig. 1). The Motor Vehicle Collisions (MVC) occurring between 2014-2023 were considered. The MVC dataset is comprised of information related to property damage collisions, fail-to-remain collisions, injury collisions, and fatalities [5]. For further details, please check the Toronto Police Service Public Safety Data Portal.

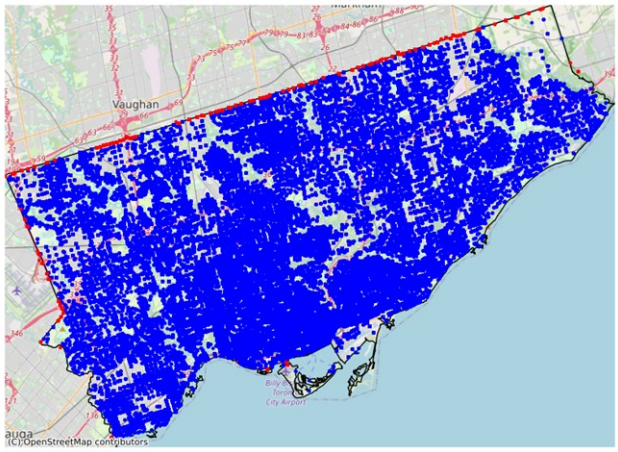

Fig. 1: The study area with the occurrences of Motor Vehicle Collisions: blue and red dots represent the occurrences inside and outside the study boundary, respectively

The dataset contained 634,858 collision records. After removing faulty records (91,791) with latitude and longitude values that fell outside the Canadian boundary, a total of 543,067 records were considered for the synthetic data generation process in CTGAN. A total of 21 variables were considered: identification codes (OBJECTID, EVENT_UNIQ); occurrence details comprising occurrence date, month, day of the week, year, and hour (OCC_DATE, OCC_MONTH, OCC_DOW, OCC_YEAR, OCC_HOUR); Police division of occurrence (DIVISION); number of persons killed or injured (FATALITIES, INJURY_COL); collisions fail to remain or property damage type (FTR_COLLIS, PD_COLLISI); neighborhood id and name (HOOD_158, NEIGHBOURH); longitude and latitude (LONG_WGS84. LAT_WGS84); information indicating whether the collision involved a person in an automobile, motorcycle, passenger, bicycle or pedestrian (AUTOMOBILE, MOTORCYCLE, PASSENGER, BICYCLE, PEDESTRIAN).

2.2 Synthesizer: CTGAN

The GAN-based deep learning synthesizer, CTGAN, available in the Synthetic Data Vault (SDV) library, was employed to train the model

and generate synthetic data [2]. The CTGAN model was trained for 75 epochs, and a batch size of 1000 was used. The values for all other parameters were set to default values, which are output samples for each one of the discriminator layers (256,256), discriminator weight decay (1e-6) and learning (2e-4) rates, size of generator's random sample (128), generator weight decay (1e-6) and learning (2e-4) rates, output samples for each one of the Residuals in the generator (256, 256). As this study mainly focused on understanding the abilities of CTGAN to generate synthetic geolocation data, default parameters were used to maintain consistency in the different synthetic data generation approaches.

2.3 Synthetic Data Generation Approaches

To demonstrate the challenges associated with modeling latitude and longitude values in CTGAN, we adopted three specific approaches:

- Approach 1 (No Constraints): The entire MVC dataset was entered into the synthesizer without applying any filtering or constraints for the latitude and longitude values. As shown in Fig1, 14,326 points were found to be located outside the City of Toronto, which were entered along with the points that fell inside.
- Approach 2 (Upper and Lower Bound Constraints): The entire MVC dataset was entered into CTGAN, excluding the 14,326 points outside the Toronto boundary. Additionally, the minimum and maximum values of latitude and longitude for the geographical extent of Toronto were computed and used as a scalar range constraint in the synthesizer.
- Approach 3 (Geospatial Aggregation): The entire MVC dataset was entered into the synthesizer, excluding the latitude and longitude values. The latitude and longitude values were used to create a Geohash variable. The Geohash is an encoding system by which latitude and longitude pairs are converted into a single Base32 string [10]. The Geohash system divides the geographic area of the whole world into rectangular grids. Therefore, the latitude and longitude values are aggregated into a rectangular grid system through the process [10]. The geohash2 library in Python was used to encode latitude and longitude.

2.4 Evaluation Techniques

The geolocation information of each of the synthetic datasets produced was evaluated in terms of their structural and spatial characteristics.

Structural/Non-spatial Evaluation Hellinger Distance (HD) measures the difference in distribution between each variable in the real and synthetic data [1, 11]. HD was implemented to quantify how different the distributions of the latitude and longitude variables were in the real and synthetic data. Given two discrete probability distributions $P = \{p_1, p_2, ..., p_n\}$ and $Q = \{q_1, q_2, ..., q_n\}$, the HD between P and Q can be expressed by eq. 1.

$$HD^2(p,q) = \frac{1}{2}\sum_{i=1}^{n}(\sqrt{p_i} - \sqrt{q_i}) \quad (1)$$

The HD scores, when close to 0, suggest that the distribution of all variables is similar between the real and the synthetic datasets.

Additionally, univariate comparisons of the latitude and longitude variables in the real and synthetic datasets were compared using histograms and Kernel Density Estimate (KDE) plots.

Spatial Evaluation Spatial distribution of events was evaluated to measure the proportion of collision events that fell within and outside the boundary of the study area. In contrast to HD metrics, this evaluation aimed to understand the actual spatial distribution of the events and can be expressed using eq. 2.

$$\text{Spatial Proportion} = \left(\frac{\text{Total events within/outside boundary}}{\text{Total of all events}}\right) \times 100 \quad (2)$$

Moreover, the spatial distribution of events was visually evaluated by producing maps of the synthetic MVC events.

3 Results and Discussions

The synthetic data generation process was successful for Approaches 1 and 2. However, the synthesizing process for Approach 3 was terminated due to excessive memory usage with the following error message.

TerminatedWorkerError: A worker process managed by the executor was unexpectedly terminated. This could be caused by a segmentation fault while calling the function or by excessive memory usage causing the Operating System to kill the worker. The exit codes of the workers are {SIGKILL(-9)}

A detailed inspection of the error message revealed that the categorical representation of the Geohash variables created an array that the CTGAN synthesizer could not handle properly. The number of unique Geohash codes in the training dataset was 14,821, which, during the one-hot encoding process for 543,067 observations, led to the memory demand to handle an array of size 543,067 x 14,821. Consequently, the synthetic data generation process using Approach 3 was terminated due to memory constraints.

Structural/Non-spatial Evaluation: The overall Hellinger Distance scores have been illustrated using the boxplot diagram in Fig. 2. The boxplots for the synthetic datasets generated using Approach 1 and 2 suggest that the distributions between real and synthetic data of all variables are very similar for synthetic data generated using Approach 1 when compared to Approach 2. Therefore, the HD results suggest that the exclusion of out-of-bounds observations in the real data and the application of upper and lower-bound constraints for the latitude and longitude values in the synthesizer could potentially affect the training and the synthetic data generation process.

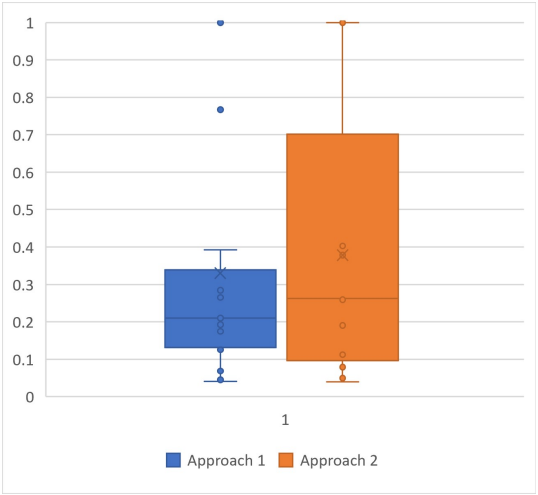

Fig. 2: The HD scores for all variables in the dataset, showing the similarities in the univariate distributions between the real and the synthetic data.

On a closer inspection of the histogram and KDE plots for the latitude and longitude values in the real and synthetic datasets (Approaches 1 and 2), the results suggest that the CTGAN had a relatively low success in retaining the distribution patterns of the latitude and longitude in the real data while producing the synthetic data Fig. 3 and Fig. 4. The individual HD scores for latitude and longitude variables were analyzed further and were found to be close to 0.9. This further confirmed that the CTGAN had difficulties in capturing the distribution of these variables from the real to synthetic datasets.

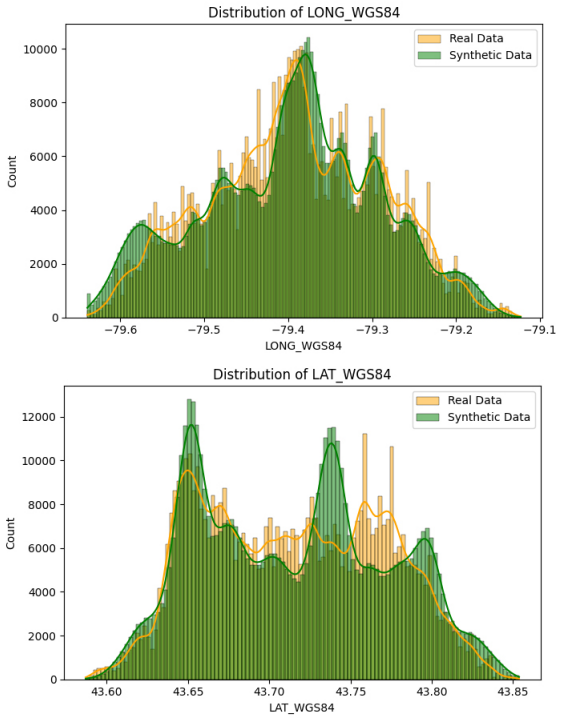

Fig. 3: The histogram and KDE plots of latitude and longitude variables in the synthetic dataset produced using Approach 1

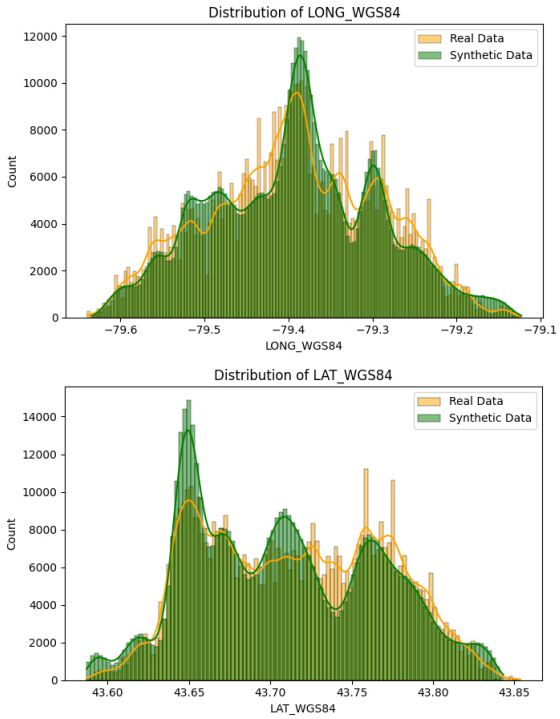

Fig. 4: The histogram and KDE plots of latitude and longitude variables in the synthetic dataset produced using Approach 2

Spatial Evaluation: Fig. 5 and Fig. 6 maps the MVC events in the synthetic datasets produced by Approach 1 and 2, respectively. The figures illustrate clearly that the latitude and longitude combinations produced in both the constrained and the unconstrained synthesizers represented a substantial number of events that were located outside the study boundary. Contrasting the two figures with Fig. 1, it is clearly visible that even for points that were located inside the study boundary, the number of unique points far exceeded the number of unique points present in the real dataset.

Table 1 shows that the number of MVC occurrences falling outside the boundary has become thrice in the synthetic datasets when compared with the real data. Moreover, the number of unique latitude and longitude pairs in the synthetic datasets is 25 times higher than the real data and is unique for each observation.

Table 1: Spatial evaluation of geolocation attributes

	Real Data	Synth. (App. 1)	Synth. (App. 2)
Unique Lat/Long	21,071	543,067	528,269
Inside Obs (%)	528,269 (97.28%)	502,481 (92.53%)	488,233 (92.42%)
Outside Obs (%)	14,798 (2.72%)	40,586 (7.47%)	40,036 (7.58%)
Total Obs	543,067	543,067	528,269

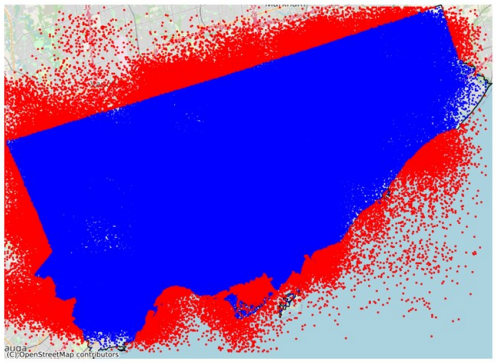

Fig. 5: The spatial distribution of MVCs in the synthetic data generated using Approach 1. The blue and red dots represent the occurrences inside and outside the study boundary, respectively

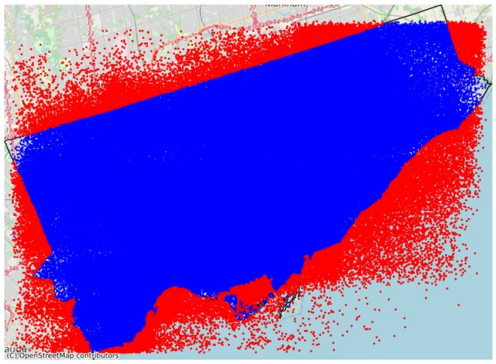

Fig. 6: The spatial distribution of MVCs in the synthetic data generated using Approach 2. The blue and red dots represent the occurrences inside and outside the study boundary, respectively

4 Conclusion

Generating synthetic datasets with geolocation information bears great potential for advancing geospatial research, software development, and the future development of deep learning models. However, the use of existing synthesizers, such as CTGAN, warrants caution as modeling latitude and longitude values in GANs requires careful consideration of the spatial characteristics of these variables. This study demonstrated that using CTGAN to generate synthetic latitude and longitude values may produce geographically redundant values. The findings form the knowledge base for future research to explore more specialized GANs for modeling geolocation data and more sophisticated techniques to model latitude and longitude values in existing GAN-based synthesizers.

Acknowledgments

We thank the Toronto Police Service for the Motor Vehicle Collision data and Statistics Canada for the boundary file of the City of Toronto.

References

[1] K. El Emam, L. Mosquera, and R. Hoptroff, *Practical synthetic data generation: balancing privacy and the broad availability of data*. O'Reilly Media, 2020.

[2] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," *Advances in neural information processing systems*, vol. 32, 2019.

[3] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative adversarial networks: An overview," *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.

[4] H. Alatrista-Salas, P. Montalvo-Garcia, M. Nunez-del Prado, and J. Salas, "Geolocated data generation and protection using generative adversarial networks," in *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, 2022, pp. 80–91.

[5] Analytics and I. of Toronto Police, "Public safety data portal: Open data documentation," Toronto Police, Technical Report, 10 2023, published October 6, 2023.

[6] J. Law and M. Quick, "Exploring links between juvenile offenders and social disorganization at a large map scale: A bayesian spatial modeling approach," *Journal of Geographical Systems*, vol. 15, pp. 89–113, 2013.

[7] R. Haining, J. Law, R. Maheswaran, T. Pearson, and P. Brindley, "Bayesian modelling of environmental risk: example using a small area ecological study of coronary heart disease mortality in relation to modelled outdoor nitrogen oxide levels," *Stochastic Environmental Research and Risk Assessment*, vol. 21, pp. 501–509, 2007.

[8] C.-H. Lin, E. Yumer, O. Wang, E. Shechtman, and S. Lucey, "St-gan: Spatial transformer generative adversarial networks for image compositing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 9455–9464.

[9] F. Zhan, H. Zhu, and S. Lu, "Spatial fusion gan for image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3653–3662.

[10] C. Zhou, H. Lu, Y. Xiang, J. Wu, and F. Wang, "Geohashtile: Vector geographic data display method based on geohash," *ISPRS International Journal of Geo-Information*, vol. 9, no. 7, p. 418, 2020.

[11] S. Gomatam, A. F. Karr, and A. P. Sanil, "Data swapping as a decision problem," *Journal of Official Statistics*, vol. 21, no. 4, p. 635, 2005.