Generative Models for Simulating Mobility Trajectories

Vaibhav Kulkarni
Department of Information Systems
UNIL-HEC Lausanne
vaibhav.kulkarni@unil.ch

Natasa Tagasovska
Department of Information Systems
UNIL-HEC Lausanne
natasa.tagasovska@unil.ch

Thibault Vatter
Department of Statistics
Columbia University
tv2233@columbia.edu

Benoit Garbinato
Department of Information Systems
UNIL-HEC Lausanne
benoit.garbinato@unil.ch

Abstract

Mobility datasets are fundamental for evaluating algorithms pertaining to geographic information systems and facilitating experimental reproducibility. But privacy implications restrict sharing such datasets, as even aggregated location-data is vulnerable to membership inference attacks. Current synthetic mobility dataset generators attempt to superficially match *a priori* modeled mobility characteristics which do not accurately reflect the real-world characteristics. Modeling human mobility to generate synthetic yet *semantically and statistically realistic* trajectories is therefore crucial for publishing trajectory datasets having satisfactory utility level *while preserving user privacy*. Specifically, long-range dependencies inherent to human mobility are challenging to capture with both discriminative and generative models. In this paper, we benchmark the performance of recurrent neural architectures (RNNs), generative adversarial networks (GANs) and nonparametric copulas to generate synthetic mobility traces. We evaluate the generated trajectories with respect to their geographic and semantic similarity, circadian rhythms, long-range dependencies, training and generation time. We also include two sample tests to assess statistical similarity between the observed and simulated distributions, and we analyze the privacy tradeoffs with respect to membership inference and location-sequence attacks.

1 Introduction

The pervasiveness of mobile devices equipped with internet connectivity and global-positioning functionality has resulted in an increasingly large amount of location-data on individuals. This data is beneficial to address and validate spatiotemporal data-based problems; predictive and kNN queries, object tracking, mobility modeling and location privacy among others. Due to the sensitive nature of datasets containing mobility traces, sharing them with untrusted entities present privacy implications. Trivial heuristics can be applied on such datasets to derive personally identifiable information of individuals, even at aggregate levels [35].

Publicly accessible mobility datasets [39, 24, 14] are usually not adequate for large scale experimental evaluations, compromising scalability tests. This issue incentivizes synthetic mobility trajectory generators that simulate the behavior of moving objects required to attain comprehensive performance valuations. In this context, one typically considers rigid and unnatural mobility models, not

Table 1: Categorization of current approaches to generate synthetic trajectories and parameters.

Technique	Model name	Parameters considered
Free movement	GSTD [34]	statistical distributions (mean, skew, standard deviation)
	G-TERD [26]	speed, rotation-angle, direction
	Oporto [8]	start time, end time, velocity, orientation
Road networks	Brinkoff [4]	speed, street capacity, nearby object location, shortest path
	SUMO [2]	road length, headway time, lane change times
	BerlinMOD [5]	road network, trip start and end, Brinkoff model
	ST-ACTS [9]	Geo-dependency model
	Hermoupolis [29]	mobility pattern, road network, points of interest
Multi environments	MWGen [36]	trip plan, road network, floor plan, routing graph
	MNTG [23]	movement model, moving objects, simulation time
Sequential models	Markov models [3]	semantic locations, geographic
	Semi-Markov models [1]	stay points, transition paths

guaranteeing the existence or even cardinality of patterns within the synthetic population. Alternative approaches rely on parametric sequential models [16] and Markov processes [3] to learn and generate trajectories. Such techniques also ignore the presence of long-range dependencies[20] inherent to human mobility which features non-Markovian character [38, 17].

It is therefore imperative to generate context-dependent synthetic traces resembling the human-mobility behavior at satisfactory utility levels while preserving user privacy. However, one of the major challenges is the absence of quantitative methods for evaluating the realistic nature of synthetic traces and the associated utility-privacy tradeoff.

To this end, we present several nonparametric approaches to generate large-scale synthetic trajectories by training the models on a real-world dataset followed by hallucinating trajectories using the trained model. We perform an extensive evaluation of the generated trajectories by assessing their geographic and semantic similarity compared to the actual dataset. We use two sample metrics to obtain the statistical similarity between datasets. We then quantify the presence of long-range dependencies by computing the mutual-information decay and conduct privacy-leakage tests on the generated trajectories. We conclude with a discussion on appropriate strategies and applicable evaluation metrics based on our experimental results and tackle open questions and challenges.

2 Related Work

Table 1 provides a summary of existing trajectory generators, where they formulate the synthetic trajectory simulation as an optimization problem, solved by genetic algorithms under the constraint of *a priori* determined parameters. A fundamental issue is the selection and definition of the parameter space that controls the evolution of the moving objects. The stringent and classified network connections thus influence the realistic nature of the generated trajectories. In several cases, there is no correlation between the future direction of movement and the past locations. Repeated visits to a given location within a short span of time are also observed due to the bounding parameters. Therefore, the symbolic nature of these frameworks result in an implicit location-dependent context, which compromises the realistic nature of the generated activity patterns. To address these drawbacks associated with parametric modeling, Ouyang et al. [28] propose a GAN-based approach to generate trajectories, where the discriminator is based on a convolutional neural network (CNN) [19]. Similarly, we explore other deep learning architectures based on RNNs known to model sequential data better than CNNs [31]. We also investigate generative models based on the nonparametric copulas of [7].

3 Synthesizing Trajectories using Generative Modeling

First we explore the benefits of applying deep learning architectures to synthesize mobility trajectories. RNNs use their hidden memory representation to process input sequences and we select four architectures: (1) Char-RNN (SRNN) [11], (2) RNN-LSTM [12], (3) recurrent highway networks (RHN) [40], and (4) pointer sentinel mixture model (PSMM) [22]. For GANs, where two neural networks compete in a zero-sum game framework, we select two architectures: (1) SGAN [37], and (2) RGAN [6]. These architectures differ in their capacity to manipulate their internal memory representation and propagate gradients along the network. In addition to neural-network based solutions, we also evaluate *copulas*; a seldom explored generative model in the machine learning
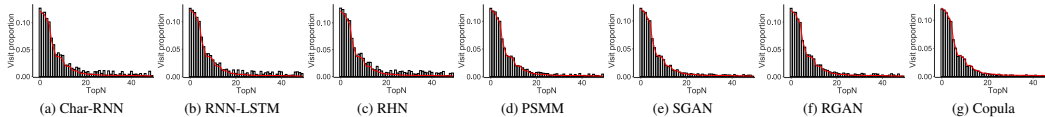
Figure 1: TopN visited locations for real and synthetic trajectories generated by each method. We select $N = 50$ out of a total of 286 locations. The red curve shows the distribution for the true dataset.

community. Given a bivariate random vector (X_1, X_2), Sklar's theorem [33] states that the joint density[1] is $f(x_1, x_2) = f_1(x_1)f_2(x_2)c(F_1(x_1), F_2(x_2))$, where f and f_i are the marginal densities, F_i the marginal distributions, and c the copula density. In other words, the bivariate density can be uniquely described by the product between its marginal densities and a copula density representing its dependence structure. A useful consequence of this representation is that, by taking the logarithm on both sides, estimation of the joint density can be performed in two steps: the marginal distributions first, and the copula afterwards. In a nutshell, copulas allow to flexibly specify the marginal and joint behavior of random variables.

An important aspect in the generative context is that, because $U = F(X) \sim U(0, 1)$ for any continuous random variable with distribution F, the copula is a distribution with uniform margins. Hence, from a copula sample (U_1, U_2), one obtains a sample on the original scale using the inverse cumulative distributions via $(X_1, X_2) = (F_1^{-1}(U_1), F_2^{-1}(U_2))$. For further details on copulas, we refer the reader to [13]. In this paper, we combine the kernel-based nonparametric copulas of [7] with the empirical distribution function of the margins obtain highly flexible models.

Data representation Given a dataset of n mobility trajectories, where a trajectory T_u of an individual u is a temporally ordered sequence of tuples, such that, $T_u = \langle(l_1, t_1), (l_2, t_2)...(l_n, t_n)\rangle$, where $l_i = (lat_i, lon_i), 0 \leq i \leq n$, the latitude-longitude coordinate pair and t, the timestamp such that $t_{i+1} > t_i$. We first transform the location data onto a uniform grid for dimensionality reduction using a technique that preserves spatial locality[2], thus translating T_u into a 2-D trace $S(t) = \langle(c_1, t_1), (c_2, t_2)...(c_n, t_n)\rangle$, where c_i is the *geo-hash* of the projected cell ID and the timestamp t_i.

4 Experiments, Results and Discussion

Experimental setup A complete trajectory sequence can be generated by iteratively feeding the current output trajectory sequence as input for the next step to the trained model. RNNs are trained on the geo-hashes and timestamps of all the individuals present in the dataset in a deterministic framework. GANs are first trained to model and then successively reproduce the traces in the same representation, which is mapped back to the (lat_i, lon_i) coordinates. We use the standard implementations of the predictive algorithms and hyper-parameters as described in their respective papers. To use copulas as generative models, we rely on the *rvinecopulib* package [25], whose `vine` routine implements the automatic kernel-based fitting of the dependence structure.

Dataset Experiments are performed using the Nokia mobile dataset [18] that consists of mobility trajectories of individuals collected in Switzerland. We use a total of 70M data points to train the considered models.

Evaluation We perform the evaluation of the generated trajectories using this dataset from four distinct dimensions: (1) geographic and semantic similarity, (2) statistical similarity (3) long-range dependencies and (4) privacy tests. In order to assess the geographic and semantic similarity, we compare the probability distribution of visiting $topN$ locations (visit-time and dwell-time) in the generated trajectories for each technique compared to the true dataset (see Figure 1). Char-RNN, RGAN and copulas have the closest fit to the true distribution indicating that the $topN$ locations are very well preserved in the respective synthetic datasets.

To evaluate the statistical similarity, we use Mean Maximum Discrepancy (MMD) [10] to test whether one can reject the null hypothesis that a synthetic sample has the same distribution as the data. MMD works by replacing the probability densities with embeddings that facilitate the computation of distances between distributions. Note that defining distance metrics in the context of time series

---

[1]It is usually stated for the distribution rather than the density and for random vectors of arbitrary dimension.
[2]Google S2: https://s2geometry.io/

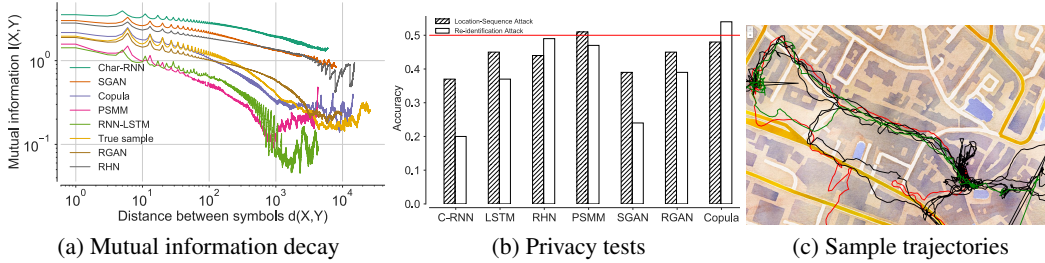| (a) Mutual information decay | (b) Privacy tests | (c) Sample trajectories |

Figure 2: (a) long-range dependency test (symbols denote individual location coordinates), (b) privacy test with location hiding as privacy preserving mechanism (red line indicates a random guess) (c) sample trajectories generated by two best approaches (copulas (black) and SGANs (red)) follow the road network for the most part and also synthesize stays at some locations indicating a point of interest. Trajectory from the actual dataset in the same area is depicted in green.

data such as mobility trajectories is challenging due to the alignment concerns [6]. We thus consider the time axis for alignment as done by Esteban et al. [6]. The results along with the training and generation time for each approach is shown in Table 2, where we observe that all approaches achieve similar results in terms of MMD, with copulas standing out with a lower value. We can thus infer that copulas can synthesize distributions with statistical characteristics closer to the observed ones. Regarding the computational efficiency, copulas require a fraction of the time needed by NN-based approaches.

Table 2: Mean and standard deviation of real vs. synthetic data (lower is better) from 30 repetitions. Second row is CPU time indicating the training/fit+generation time.

Metric/Method	Char-RNN	RNN-LSTM	RHN	PSMM	SGAN	RGAN	Copula
MMD	0.32(1e-3)	0.27(9e-4))	0.30(1e-3)	0.21(6e-4)	0.19(7e-4)	0.21(6e-4)	0.01(6e-4)
CPU time (sec)	9k+∼10	10.3k+∼14	12.7k+∼15	10.5k+∼15	11.2k+∼15	11.5k+∼14	6.5 + 0.76

Figure 2(a) shows the result of long-range dependency test, in terms of mutual information decay [21, 20]. We observe a power-law decay in case of GANs, copulas and RNN-LSTM indicating that they account for the long-range dependencies in mobility trajectories. Figure 2(b) shows the results of two privacy tests: (1) location-sequence attack, and (2) membership interference attack. Given a synthetic dataset, (1) answers to what level of accuracy can trajectories in the dataset be reconstructed [32], and (2) an adversary's accuracy of inferring if a target individual contributed to the specific trajectory [30]. For these tests, we use the the location-privacy and mobility meter [32], where obfuscation is performed using the location hiding mechanism. Given a completely random distribution the accuracy of a recovered user-information is 0, we therefore suspect that the privacy-based score is biased towards representations which do not accurately capture the statistical properties of the true dataset.

5   Conclusion and Future Work

In this work, we propose and evaluate a variety of generative models to synthesize mobility trajectories. To the best of our knowledge, this is the first study to do so using seven different approaches while evaluating their realism across four dimensions. From the results and discussion, we observe that regarding statistical and semantic properties, copulas have an advantage over all other methods. Additionally, all NN-based methods are time consuming, which makes copulas favorable when computational efficiency is important to the end-users. As future work, we will consider datasets collected in bigger cities and generate larger synthetic datasets to evaluate the performance of these models under high movement stochasticity. From curbing the privacy leakage of the true dataset while maintaining utility, trajectory generation can be designed as an optimization problem with an objective to jointly maximize statistical similarity and privacy. But it is still not clear how to assess such property and adaptive/configurable metric, and is part of ongoing work [27]. While this paper represents an initial comparative study of various generative models, a deeper understanding of their performances will be needed to compute utility-privacy scores as applied to online services by Krause and Horvitz [15] before publicly releasing the synthetic datasets. Another interesting avenue for research is to apply transfer learning in order to map a mobility behavioral model captured in one city on to another region.

References

[1] Mitra Baratchi, Nirvana Meratnia, Paul JM Havinga, Andrew K Skidmore, and Bert AKG Toxopeus. A hierarchical hidden semi-markov model for modeling mobility data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 401–412. ACM, 2014.

[2] Michael Behrisch, Laura Bieker, Jakob Erdmann, and Daniel Krajzewicz. Sumo–simulation of urban mobility. In *The Third International Conference on Advances in System Simulation (SIMUL 2011), Barcelona, Spain*, volume 42, 2011.

[3] Vincent Bindschaedler and Reza Shokri. Synthesizing plausible privacy-preserving location traces. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 546–563. IEEE, 2016.

[4] Thomas Brinkhoff. A framework for generating network-based moving objects. *GeoInformatica*, 6(2):153–180, 2002.

[5] Christian Düntgen, Thomas Behr, and Ralf Hartmut Güting. Berlinmod: a benchmark for moving object databases. *The VLDB Journal—The International Journal on Very Large Data Bases*, 18(6):1335–1368, 2009.

[6] Cristóbal Esteban, Stephanie L Hyland, and Gunnar Rätsch. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633*, 2017.

[7] Gery Geenens, Arthur Charpentier, and Davy Paindaveine. Probit transformation for non-parametric kernel estimation of the copula density. *Bernoulli*, 23(3):1848–1873, 2017. ISSN 1350-7265.

[8] Fosca Giannotti, Andrea Mazzoni, Simone Puntoni, and Chiara Renso. Synthetic generation of cellular network positioning data. In *Proceedings of the 13th annual ACM international workshop on Geographic information systems*, pages 12–20. ACM, 2005.

[9] Gyozo Gidofalvi and Torben Bach Pedersen. St–acts: a spatio-temporal activity simulator. In *Proceedings of the 14th annual ACM international symposium on Advances in geographic information systems*, pages 155–162. ACM, 2006.

[10] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[11] Stephen Grossberg. Recurrent neural networks. *Scholarpedia*, 8(2):1888, 2013.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 8: 1735–80, 1997.

[13] Harry Joe. *Dependence Modeling with Copulas*. Chapman & Hall/CRC, 2014.

[14] Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin*, 2010.

[15] Andreas Krause and Eric Horvitz. A utility-theoretic approach to privacy and personalization. In *AAAI*, volume 8, pages 1181–1188, 2008.

[16] Vaibhav Kulkarni and Benoît Garbinato. Generating synthetic mobility traffic using rnns. In *Proceedings of the 1st Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery*, GeoAI '17, pages 1–4, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5498-1. doi: 10.1145/3149808.3149809. URL http://doi.acm.org/10.1145/3149808.3149809.

[17] Vaibhav Kulkarni, Abhijit Mahalunkar, Benoit Garbinato, and John D Kelleher. On the inability of markov models to capture criticality in human mobility. *arXiv preprint arXiv:1807.11386*, 2018.

[18] Juha K Laurila, Daniel Gatica-Perez, Imad Aad, Olivier Bornet, Trinh-Minh-Tri Do, Olivier Dousse, Julien Eberle, Markus Miettinen, et al. The mobile data challenge: Big data for mobile computing research. In *Pervasive Computing*, number EPFL-CONF-192489, 2012.

[19] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[20] Henry W Lin and Max Tegmark. Critical behavior from deep dynamics: a hidden dimension in natural language. *arXiv preprint arXiv:1606.06737*, 2016.

[21] Abhijit Mahalunkar and John D Kelleher. Using regular languages to explore the representational capacity of recurrent neural architectures. *arXiv preprint arXiv:1808.05128*, 2018.

[22] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *CoRR*, abs/1609.07843, 2016.

[23] Mohamed F Mokbel and Walid G Aref. Sole: scalable on-line execution of continuous queries on spatio-temporal data streams. *The VLDB Journal—The International Journal on Very Large Data Bases*, 17(5):971–995, 2008.

[24] Sonia Ben Mokhtar, Antoine Boutet, Louafi Bouzouina, Patrick Bonnel, Olivier Brette, Lionel Brunie, Mathieu Cunche, Stephane D'Alu, Vincent Primault, Patrice Raveneau, et al. Priva'mov: Analysing human mobility through multi-sensor datasets. In *NetMob 2017*, 2017.

[25] Thomas Nagler and Thibault Vatter. rvinecopulib: high performance algorithms for vine copula modeling, 2018. URL `https://cran.r-project.org/package=rvinecopulib`.

[26] Mario A. Nascimento, Dieter Pfoser, and Yannis Theodoridis. Synthetic and real spatiotemporal datasets. *IEEE Data Eng. Bull.*, 26(2):26–32, 2003.

[27] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. *arXiv preprint arXiv:1807.05852*, 2018.

[28] Kun Ouyang, Reza Shokri, David S Rosenblum, and Wenzhuo Yang. A non-parametric generative model for human trajectories. In *IJCAI*, pages 3812–3817, 2018.

[29] Nikos Pelekis, Stylianos Sideridis, Panagiotis Tampakis, and Yannis Theodoridis. Hermoupolis: a semantic trajectory generator in the data science era. *SIGSPATIAL Special*, 7(1):19–26, 2015.

[30] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. Knock knock, who's there? membership inference on aggregate location data. *arXiv preprint arXiv:1708.06145*, 2017.

[31] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61: 85–117, 2015.

[32] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. Quantifying location privacy. In *2011 IEEE symposium on security and privacy*, pages 247–262. IEEE, 2011.

[33] 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.

[34] Yannis Theodoridis, Jefferson RO Silva, and Mario A Nascimento. On the generation of spatiotemporal datasets. In *International Symposium on Spatial Databases*, pages 147–164. Springer, 1999.

[35] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1241–1250. International World Wide Web Conferences Steering Committee, 2017.

[36] Jianqiu Xu and Ralf Hartmut Güting. Mwgen: a mini world generator. In *Mobile Data Management (MDM), 2012 IEEE 13th International Conference on*, pages 258–267. IEEE, 2012.

[37] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pages 2852–2858, 2017.

[38] Zhi-Dan Zhao, Shi-Min Cai, and Yang Lu. Non-markovian character in human mobility: Online and offline. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(6):063106, 2015.

[39] Yu Zheng, Xing Xie, and Wei-Ying Ma. Geolife: A collaborative social networking service among user, location and trajectory. *IEEE Data Eng. Bull.*, 33:32–39, 2010.

[40] Julian Georg Zilly, Rupesh Kumar Srivastava, Jan Koutník, and Jürgen Schmidhuber. Recurrent highway networks. In *ICML*, 2017.