Generating Synthetic Data on Individual Daily Trajectories with an Autoregressive Language Model

Keywords: Human mobility, Trajectory generation, Spatiotemporal model, Transformer, NLP

Extended Abstract

The acquisition of big data on individual daily trajectories is crucial in addressing issues related to disasters, terrorism, public safety, infectious diseases, spatial segregation, marketing, and traffic congestion. By analyzing the big data, we can detect the causes of traffic congestion and monitor the evacuation of people in natural disasters [1]. However, the use of big data is not without challenges, particularly with regard to safeguarding individual geo-privacy. It is difficult to control the trade-off between uncertainty and utility when disclosing real-world data on human mobility. Synthetic trajectories that preserve statistical properties have the potential to achieve performance comparable to real-world data on multiple tasks to solve social issues.

Both physics and machine learning approaches have been adopted to develop trajectory generation models [2, 3]. In this study, we propose a model for generating individual daily trajectories using GPT-2 [4], which is one of the Transformer models that has emerged as a successful alternative to Recurrent Neural Network in natural language generation.

In order to train our trajectory generation model, we utilized location data obtained from Agoop Corp., comprising 128.46 million logs collected from a total of 670,000 smartphones (about 22,000 per day) passing through Urayasu City, Chiba, Japan during August 2022. Notably, Urayasu City is the site of Tokyo Disney Resort, a globally renowned theme park that attracts over 30 million visitors annually. With its convenient access to central Tokyo, the city is only a 13-minute train ride away. The location data includes latitude, longitude, and timestamps of departure and arrival at each location, as well as the unique ID of each smartphone owner, their gender, age, and home location. The time resolution of this data is on average 7 minutes. In this study, we model the trajectories of smartphone owners when they are not at home.

We aim to generate individual daily trajectories applying the autoregressive language model, "GPT-2," developed by OpenAI [4]. GPT-2 consists of multiple Transformer layers composed of Self-attention and Projection layers, similar to BERT. GPT-2 can sequentially predict the next token from the previous token, i.e., the next location from the past locations, by referring only to the input token sequence before the position to be processed in the Transformer layers.

A trajectory can be represented as a time series of travel time intervals and origin-destination coordinates. However, language models like GPT-2 are not well-suited to learning numerical values such as time and coordinates. As a result, we opted to convert these values into unique characters. To accomplish this, we began by discretizing the time interval Δt, with $\tau = int(\log \Delta t)$, and assigned a unique character to each resulting discrete time interval τ.

$$r(\tau) \in \{R|\text{Unique characters assigned to discretized time intervals}\}.$$

Next, applying the Japanese regional grid code JIS X 0410, we transform the coordinates expressed in latitude and longitude into unique characters that recursively subdivide space. As an instance, "$\zeta_1\zeta_2\zeta_3\zeta_4\zeta_5$" represents the area at a 250m resolution. Here, the character variable ζ_i corresponds to the *i*-th level grid code of JIS X 0410. ζ_1 denotes a unique location enclosed

by a square with a 40-minute difference in latitude and a 1-degree difference in longitude. In Japan, the land areas can be represented by 176 first-level grid codes, which cover the entire country. ζ_2 designates the area created by dividing the first-level grid into eight equal areas in direction of latitude and longitude, respectively. ζ_3 also represents the area created by dividing the second-level grid into ten equal areas in direction of latitude and longitude, respectively. Subsequent divisions are recursively split into two equal areas, and each ζ_i is assigned a unique character.

$$\zeta_i \in \{\mathrm{Z}_i | \text{Unique characters assigned to areas}\},$$

where $R \cap \mathrm{Z}_i = \emptyset$ and $\mathrm{Z}_i \cap \mathrm{Z}_j = \emptyset \ \ (i \neq j)$. Land areas in Japan, at a resolution of 250 m, can be represented by combinations of only 348 $(= 176 + 8^2 + 10^2 + 2^2 + 2^2)$ characters.

An individual daily trajectory can be represented by the characters r and X assigned to the time intervals and origin-destination coordinates, respectively, as follows,

$$X(t_0)\_r(\tau_1)X(t_0 + \Delta t_1)\_r(\tau_2)X\left(t_0 + \sum\nolimits_{k=1}^{2} \Delta t_k\right)\_\cdots$$

where t_0 denotes the time of first leaving home on a given day, and Δt_k and τ_k represent the time interval and the discretized time interval from the $(k-1)$-th destination to the k-th destination, respectively. The characters X representing the area are $X(t) = \zeta_1(t)\zeta_2(t)\zeta_3(t)\zeta_4(t)\zeta_5(t)$ for 250m resolution. We insert a comma token "," for temporary returns home and a period token "." for the last return home each day. The previous and next locations are connected with a "_" to represent the trajectory. We train GPT-2 to learn individual daily trajectories expressed in this textual form.

Figure 1 depicts the trajectories generated by GPT-2 when the initial coordinates are set to the entrance gate of Tokyo Disneyland in Urayasu City. These generated trajectories replicate the realistic behavior of visitors who spend approximately 10 hours in the theme park before returning to the entrance gate. Notably, GPT-2-generated individual trajectories exhibit the following seven realistic properties: (1) the cumulative distribution of the hourly moving distance follows a logarithmic function; (2) the autocorrelation function of the moving distance exhibits short-time memory; (3) there is a positive autocorrelation in the direction of moving for one hour in long-distance trips; (4) the last location is often near the initial location in each individual daily trajectory; (5) the diffusion of people depends on the time scale of their moving; (6) the distribution of moving time intervals has a long tail; and (7) the correlation between the moving time interval and the moving distance observed in the real world is reproduced.

We assess the predictive performance of GPT-2, as well as 1st and 2nd order Markov chains, and Catboost models, for individual daily trajectories. For each model, we input the initial five coordinates of a real individual daily trajectory and predicted (generated) the coordinates for the next half hour, one hour, two hours, dour hours, and the final time (i.e., homecoming time). Table 1 shows the probability of the predicted location is within 1 km (10 km) of the actual location coordinates. Notably, GPT-2 demonstrated superior performance compared to the other models for all predictions. In particular, for the last location of the day, we observed that GPT-2 was eight times more accurate than the other models.

References

[1] Zhu L, et al. (2019) IEEE Trans Intell Transp Syst 20, 383–98.
[2] Schlapfer M, et al. (2021) Nature 593, 522–527.
[3] Mizuno T, Fujimoto S, Ishikawa A. (2022) Front. Phys. 10, 1021176.
[4] Radford A, Wu J, Child R, Amodei D, Sutskever I. (2019) OpenAI blog 1, 9.

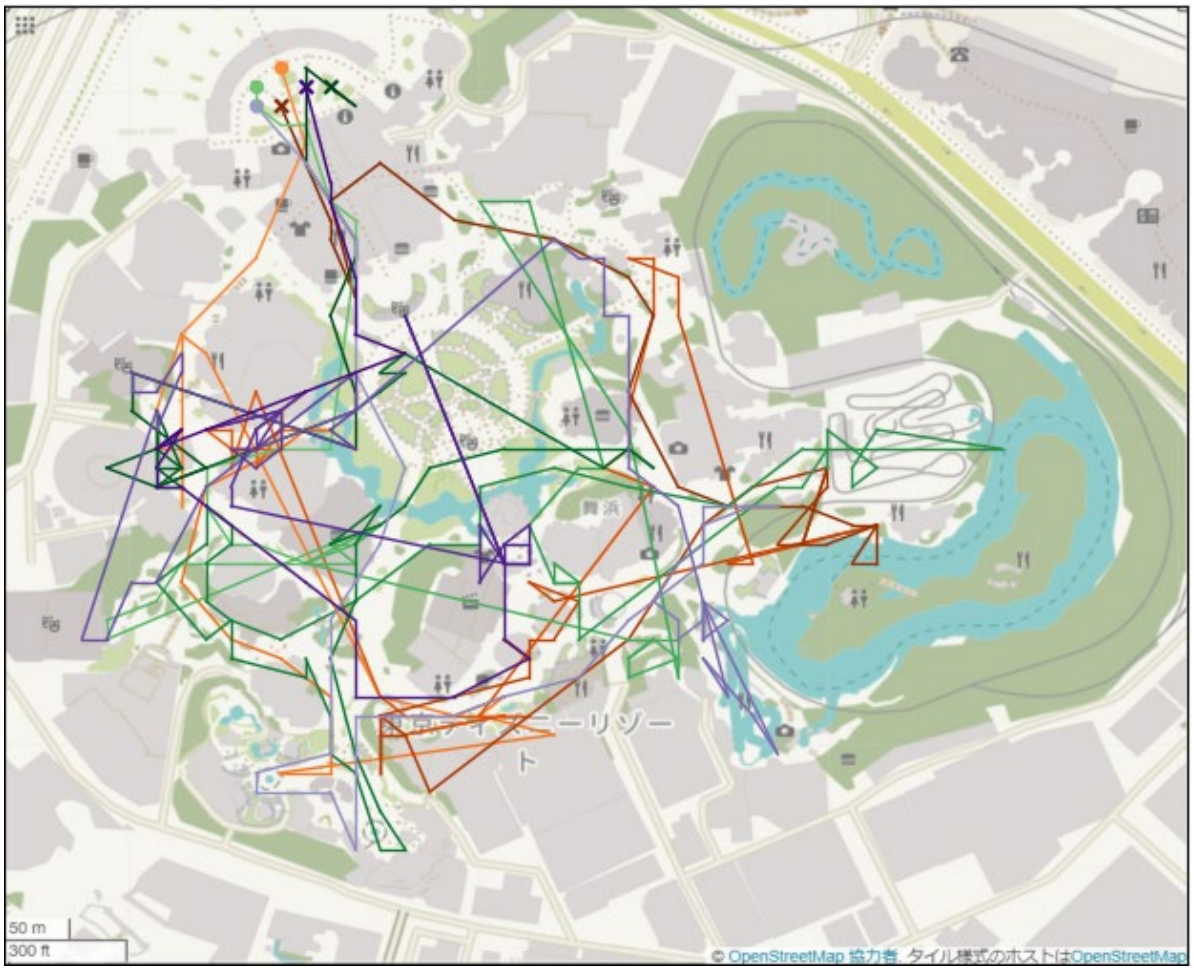

Figure 1 GPT-2 generated trajectories in Tokyo Disneyland. The initial coordinates set to the entrance gate. ● is initial location and × is final location for each trajectory.

Table 1 Probability that the prediction is within 1 km (10 km) of the actual location coordinates for the next half hour, one hour, two hours, four hours, and the final time of the day. We performed 24,247 realizations of each model to estimate the probabilities.

	30 minutes	1 hour	2 hours	4 hours	Final time of day
1st Markov	0.29 (0.67)	0.16 (0.54)	0.07 (0.42)	0.04 (0.33)	0.01 (0.20)
2nd Markov	0.33 (0.75)	0.20 (0.61)	0.11 (0.46)	0.05 (0.33)	0.01 (0.19)
Catboost	0.15 (0.70)	0.07 (0.54)	0.04 (0.45)	0.03 (0.40)	0.02 (0.25)
GPT-2	0.40 (0.82)	0.26 (0.70)	0.16 (0.55)	0.10 (0.43)	0.12 (0.40)