Urban Human Mobility Prediction Using Support Vector Regression: A Classical Data-Driven Approach

Yuki Imai
yuki.imai@ariseanalytics.com
ARISE analytics.Inc
Shibuya, Tokyo, Japan

Takuya Tokumoto
takuya.tokumoto@ariseanalytics.com
ARISE analytics.Inc
Shibuya Tokyo, Japan

Kohei Koyama
kohei.koyama@ariseanalytics.com
ARISE analytics.Inc
Chiyoda Tokyo, Japan

Tomoko Ochi
to-ochi@kddi.com
KDDI CORPORATION
Shibuya Tokyo, Japan

Shogo Imai
shogo.imai@ariseanalytics.com
ARISE analytics.Inc
Shibuya Tokyo, Japan

Tomoyuki Mori
tomoyuki.mori@ariseanalytics.com
ARISE analytics.Inc
Shibuya Tokyo, Japan

Tomohiro Nakao
tomohiro.nakao@ariseanalytics.com
ARISE analytics.Inc
Shibuya Tokyo, Japan

Kenta Maruyama
kenta.maruyama@ariseanalytics.com
ARISE analytics.Inc
Shibuya Tokyo, Japan

ABSTRACT

This paper presents an efficient method for predicting human mobility trajectories in urban areas of Japan, developed for the HuMobChallenge2024. Utilizing large-scale human mobility data, we constructed personalized models for individual users, enhancing trajectory prediction accuracy.

In the preprocessing phase, we applied linear interpolation to fill missing values and ensure the continuity of natural movement patterns. Feature engineering introduced novel mobility-related features alongside time-based variables. Among the evaluated models, Support Vector Regression (SVR) with a nonlinear kernel achieved the highest accuracy, demonstrating a 46% improvement in DTW scores over rule-based models. Although there was a slight decrease in GEO-BLEU scores, our approach improved spatial accuracy, striking a balance between computational efficiency and predictive performance.

Our findings suggest that human mobility prediction can balance computational efficiency and predictive accuracy by combining appropriate feature engineering with traditional machine learning methods, without the need for complex Transformer models. These results could have practical applications in fields such as urban planning and disaster risk management.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning approaches**; • **Human-centered computing** → **Ubiquitous and mobile computing**.

HuMob'24, October 29-November 1 2024, Atlanta, GA, USA

ACM ISBN 979-8-4007-1150-3/24/10
https://doi.org/10.1145/3681771.3699916

KEYWORDS

Human mobility, Trajectory predictions, Machine Learning, Support Vector Regression

ACM Reference Format:
Yuki Imai, Takuya Tokumoto, Kohei Koyama, Tomoko Ochi, Shogo Imai, Tomoyuki Mori, Tomohiro Nakao, and Kenta Maruyama. 2024. Urban Human Mobility Prediction Using Support Vector Regression: A Classical Data-Driven Approach. In *2nd ACM SIGSPATIAL International Workshop on the Human Mobility Prediction Challenge (HuMob'24), October 29-November 1 2024, Atlanta, GA, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3681771.3699916

1 INTRODUCTION

Accurate prediction of human mobility is essential for applications like traffic management, disaster risk reduction, and urban planning. The availability of large-scale mobility datasets, typically collected from mobile devices and social media, has facilitated the development of models capable of capturing complex movement patterns in urban environments.

Within this context, the HuMobChallenge2024 [2] provided an opportunity to explore methods for predicting human mobility patterns in Japanese metropolitan areas using a dataset provided by LY Corporation. Notably, in the previous year's HuMobChallenge2023, many top-performing teams employed Transformer-based models to achieve state-of-the-art results [3] [6] [9]. The success of these models has solidified their prominence in mobility prediction, a trend that is expected to continue as machine learning technologies advance.

Despite the high accuracy of Transformer-based models, they present notable challenges, such as high computational demands and limited interpretability.

We hypothesized that human mobility is governed by stable individual-level patterns that can be effectively captured using simpler, more interpretable models. Therefore, we adopted a table-based approach and developed personalized models for each user.

Although our method is more straightforward than Transformer-based approaches, it achieves a balance between accuracy and

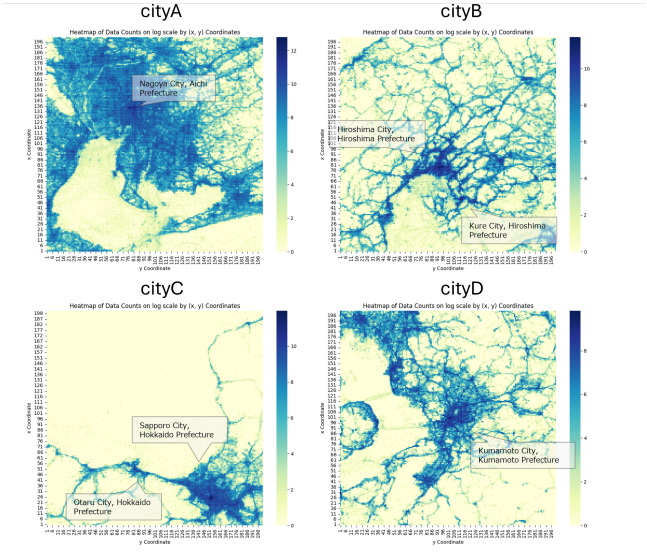


Figure 1: Human Mobility Visualization for Each City in the Dataset

computational efficiency. Our results show that by focusing on individualized patterns, even simpler models can deliver competitive performance, especially when considering cost and resource constraints.

2 DATASET DESCRIPTION

The dataset [8] contains data from four regions: A, B, C, and D. Based on the distribution of logs and file names, we inferred that each region corresponds to specific cities: Nagoya City (Region A), Hiroshima City (Region B), Sapporo City (Region C), and Kumamoto City (Region D). In this study, we focus on predicting the mobility of a subset of users (a total of 3,000 users) from datasets B, C, and D for the period from day 61 to day 75.

The target areas are divided into 500m x 500m grid cells, forming a 200 x 200 grid. Individual movements are recorded at 30-minute intervals and are discretized into the corresponding 500m grid cells. Figure 1 shows the visualization of human mobility in each city, where areas with higher mobility are clearly visible. Figure 2 shows the daily record counts in the HuMob dataset. A clear weekly pattern is visible, with a noticeable drop in logs on public holidays, helping estimate specific dates and days of the week.

Details on the data collection methods and frequency are described in Reference [8]. Each GPS record includes a unique user ID, observation time, longitude, and latitude.The data collection frequency is adjusted to minimize battery load on users' smartphones, with the logging frequency varying according to the user's movement speed. For instance, when a user remains in one location for an extended period, the logging frequency decreases, while it increases during active movement. Consequently, in City B, only about 30% of the expected logs are recorded when movement is minimal, with similar trends observed in City C and City D.

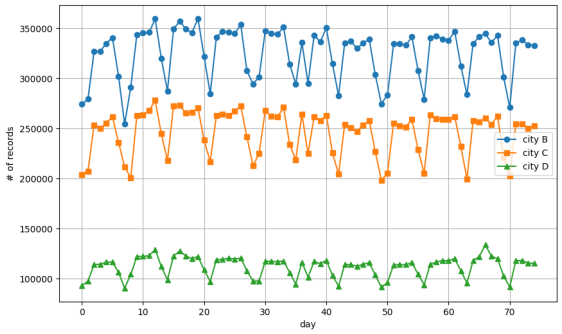


Figure 2: Daily Changes in Data Collection Logs for Each City

3 METHOD

3.1 Pre-processing

The dataset contains substantial missing data, with frequent gaps in the recorded location information. Using raw data with these gaps could result in unrealistic representations of movement patterns. To mitigate this, we applied a linear interpolation technique to estimate missing locations and ensure smoother transitions between recorded points.

Linear interpolation was used to impute missing values between the last known and next available locations when gaps extended over N consecutive hours. After evaluating multiple values of N, we determined that setting N between 8 and 12 hours yielded the highest improvement in predictive accuracy.A larger N assumes overly smooth movements, while a smaller N introduces noise due to excessive interpolation.

3.2 Feature Design

Table 1 presents the list of features used in this study. Among these, time-related features were incorporated based on the methodology outlined in reference [7].

We also introduced novel mobility-related features, hypothesizing that combining time-based and mobility-related variables would enable the model to capture individual movement patterns more effectively. While these features enhanced the model, further refinement is necessary to optimize their contribution to predictive accuracy.

In the reference that inspired the time-related features, Point of Interest (POI) visit frequencies by category, as well as cluster numbers derived from clustering features, were included. However, we did not adopt these features in our study.

Incorporating Point of Interest (POI) visit frequencies would have greatly increased the dimensionality of the dataset, leading to a substantial rise in computational cost. Given our priority to maintain a lightweight and efficient model, we opted to exclude POI features in this study.However, if computational resources and time allow, incorporating these features may further improve model accuracy.

As for the cluster numbers derived from clustering features, we chose not to include them because our approach focuses on building

Table 1: Features Introduced in the Model

Features	Reference
Date and time	[7]
Activity Time	[7]
Day of the week	[7]
Weekday or holiday	[7]
AM/PM	[7]
Number of movements	proposed
Average travel distance	proposed
Standard deviation of travel distance	proposed
Average travel angle	proposed

models at the individual level, making it challenging to account for collective movement patterns. Thus, we considered that cluster numbers would not effectively contribute to improving accuracy in this context.

3.3 Models

In this study, we compared the performance of several models, including a rule-based model (see Appendix), LightGBM and SVR, using the features described above.

Finally, the evaluation using the training and validation data showed that the SVR model with a non-linear kernel achieved the highest accuracy. Specifically, the SVR with an RBF kernel indicated strong performance in predicting human mobility. This result is consistent with findings in the literature, indicating that SVR is well-suited for this dataset and, more broadly, for human mobility prediction.

4 EXPERIMENTAL RESULTS

4.1 Evaluation Metrics

To evaluate the performance of the proposed models, we utilized two primary metrics: Dynamic Time Warping (DTW) [4] and GEO-BLEU [5]. These metrics provide complementary insights into how well the model captures both the spatial and temporal aspects of the predicted trajectories.

GEO-BLEU: GEO-BLEU is an adaptation of the BLEU score, commonly used in natural language processing, tailored for geospatial data. This metric measures the accuracy of location-based predictions by assessing the similarity between predicted and actual locations. A higher GEO-BLEU score suggests greater spatial accuracy and better alignment between the predicted and true trajectories.

Dynamic Time Warping (DTW): DTW is a widely used similarity metric for time-series data, especially when comparing sequences of different lengths or timing. It calculates the optimal alignment between two trajectories by allowing flexible adjustments in timing and speed. A lower DTW score indicates a higher similarity between the predicted and actual sequences, reflecting more accurate temporal predictions.

Using both metrics together allows us to assess the model's performance in terms of both trajectory shape and spatial alignment. DTW evaluates how well the model captures the overall sequence and timing of movements, while GEO-BLEU focuses on local spatial accuracy, providing a comprehensive evaluation of the model's predictive capabilities.

4.2 Training and Evaluation Datasets

Figure 3 shows the division of the dataset into training, evaluation, and test data. The training data consists of 2,000 users (uid = 20000 to 21999) in City B, covering a period from day 0 to day 59. The evaluation data (ground truth) comes from the same users, spanning from day 60 to day 74. It is important to note that since our approach builds individual models for each user, the models are constructed separately for each uid.

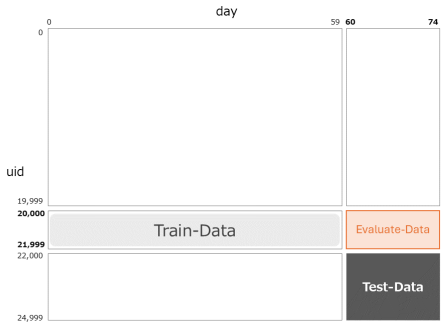


Figure 3: Segmentation of the Dataset into Training, Evaluation, and Test Sets

4.3 Evaluation Results

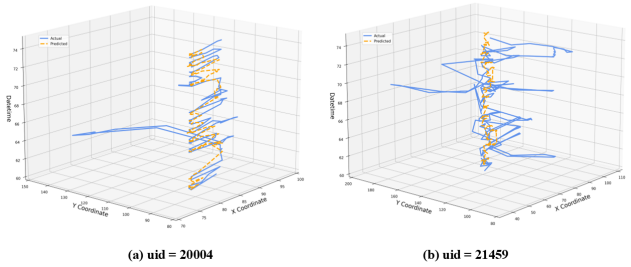
Table 2 shows the evaluation results for each method. The rule-based model achieved relatively good accuracy in terms of GEO-BLEU (0.25), while the SVR model showed a slight decrease in GEO-BLEU (0.228). However, for DTW, the SVR model demonstrated a 46% improvement (from 45.61 to 24.66).

When comparing the SVR models, there was no significant change in DTW values across models that incorporated preprocessing or mobility-related features in addition to time-related features (from 24.66 to 24.50). On the other hand, GEO-BLEU slightly improved compared to the model that used only time-related features, indicating a minor increase in spatial accuracy by incorporating mobility-related features.

Figure 4 visualizes the movement trajectories of specific users, where the solid blue line represents the actual trajectories, and the dashed orange line shows the predicted trajectories by the model. (a) For uid = 20004, a regular movement pattern is observed, and the model seems to capture this pattern to a certain extent. However, the model fails to predict the irregular movements that occurred between days 62 and 64. (b) For uid = 21459, no clear regular pattern is observed in the actual movements, which explains why the model does not accurately track the trajectory. However, the central location of movement (likely assumed to be the user's home) appears to be captured by the model.

Table 2: Evaluation Results on CityB Dataset

Method	GEO-BLEU	DTW
Rule-based	0.251	45.61
Rule-based(pre-processing)	0.265	38.59
SVR (time features)	0.228	24.66
SVR (time features + pre-processing)	0.231	24.75
SVR (time + mobility features + pre-processing)	0.231	24.50



(a) uid = 20004 (b) uid = 21459

Figure 4: Comparison of Actual and Predicted 3D Trajectories

5 CONCLUSION

This study sought to enhance the accuracy of human mobility predictions for the HuMob2024 Challenge by focusing on relatively simple models, including rule-based approaches and SVR.

The results indicate that predictive accuracy was improved not only by leveraging established features but also through preprocessing techniques and the introduction of novel mobility-related features. While the models performed well in predicting regular mobility patterns, challenges persist in capturing irregular movements.

For future work, we plan to incorporate additional factors such as POI data and event dates to improve predictions for irregular mobility patterns [1]. Introducing ensemble models could further enhance overall performance by combining the strengths of multiple algorithms. Additionally, leveraging multi-city datasets will be a crucial next step. For instance, understanding how nationwide weather patterns or holidays affect movement in other cities could improve the generalizability of the models.

Transformer-based approaches were also explored; however, stability issues were encountered in the current implementation. Despite this, Transformers remain a promising direction for mobility prediction, and we intend to refine this approach in future research.

ACKNOWLEDGMENTS

This research was conducted with the support of our company. We would like to acknowledge our company for providing the computing environment necessary for our participation in this competition.

REFERENCES

[1] Chen Cheng, Haiqin Yang, Michael R. Lyu, and Irwin King. 2013. Where you like to go next: successive point-of-interest recommendation. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (Beijing, China) *(IJCAI '13)*. AAAI Press, 2605–2611.

[2] HuMob Challenge 2024 [n. d.]. HuMob Challenge 2024. https://wp.nyu.edu/humobchallenge2024/

[3] Akihiro Kobayashi, Naoto Takeda, Yudai Yamazaki, and Daisuke Kamisaka. 2023. Modeling and generating human mobility trajectories using transformer with day encoding. In *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge* (Hamburg, Germany) *(HuMob-Challenge '23)*. Association for Computing Machinery, New York, NY, USA, 7–10. https://doi.org/10.1145/3615894.3628506

[4] Pavel Senin. 2008. Dynamic Time Warping Algorithm Review. https://api.semanticscholar.org/CorpusID:16629907

[5] Toru Shimizu, Kota Tsubouchi, and Takahiro Yabe. 2021. GEO-BLEU: Similarity Measure for Geospatial Sequences. *CoRR* abs/2112.07144 (2021). arXiv:2112.07144 https://arxiv.org/abs/2112.07144

[6] Aivin V. Solatorio. 2023. GeoFormer: Predicting Human Mobility using Generative Pre-trained Transformer (GPT). In *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge* (Hamburg, Germany) *(HuMob-Challenge '23)*. Association for Computing Machinery, New York, NY, USA, 11–15. https://doi.org/10.1145/3615894.3628499

[7] Masahiro Suzuki, Shomu Furuta, and Yusuke Fukazawa. 2023. Personalized human mobility prediction for HuMob challenge. In *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge* (Hamburg, Germany) *(HuMob-Challenge '23)*. Association for Computing Machinery, New York, NY, USA, 22–25. https://doi.org/10.1145/3615894.3628501

[8] Toru Shimizu Yoshihide Sekimoto Kaoru Sezaki Esteban Moro Alex Pentland Takahiro Yabe, Kota Tsubouchi. 2024. YJMob100K: City-scale and longitudinal dataset of anonymized human mobility trajectories. *Scientific Data* 11, 397 (April 2024), 36–44. https://doi.org/10.1038/s41597-024-03237-9

[9] Haru Terashima, Naoki Tamura, Kazuyuki Shoji, Shin Katayama, Kenta Urano, Takuro Yonezawa, and Nobuo Kawaguchi. 2023. Human Mobility Prediction Challenge: Next Location Prediction using Spatiotemporal BERT. In *Proceedings of the 1st International Workshop on the Human Mobility Prediction Challenge, HuMob-Challenge 2023, Hamburg, Germany, 13 November 2023*. ACM, 1–6. https://doi.org/10.1145/3615894.3628498

A RULE-BASED MODEL

The rule-based model operates on straightforward logic and is computationally efficient for making predictions. Although it was originally developed as a baseline for comparison with more complex machine learning models, including Transformers, the rule-based model demonstrated surprisingly competitive performance. Therefore, we have included it in our analysis.

A.1 Method

We hypothesized that each user's mobility patterns vary predictably based on the day of the week and time of day. To capture these patterns, we calculated the most frequently visited mesh codes across different temporal intervals from the dataset (0 <= d < 60). Specifically, we aggregated the most visited locations based on the following criteria:

- `Day of the week and time slot`

- `Time slot only`
- `Day of the week for daytime and nighttime`
- `Entire daytime and nighttime period`
- `Across the full time span`

Using this prioritized data, we predicted mobility for d >= 60 by extrapolating from earlier patterns.

A.2 Result

The rule-based prediction method was tested on a portion of the dataset, and we found that it achieved reasonable accuracy. As described above,the specfic results confirmed this, supporting the validity of our initial hypothesis.

While the aggregation rules in this rule-based model were manually designed, we believe that machine learning models could help automatically derive more optimal rules with lower cost. Therefore, we extended this approach by applying machine learning models based on the same hypothesis.