Clustering of Human Movement Trajectories based on Distributional Representations Derived from Bi-directional LSTM Network with Geographical Coordinates

Hiroki Tanaka, Takeshi Saga, Satoshi Nakamura
Center for Advanced Intelligence Project, RIKEN
Nara Institute of Science and Technology, Japan

IEEE BigData 2021 workshop Applications of Big Data Technology in the Transport Industry
2021/11/22

Nakamura SRG
Augmented Human Communication (AHC) Laboratory

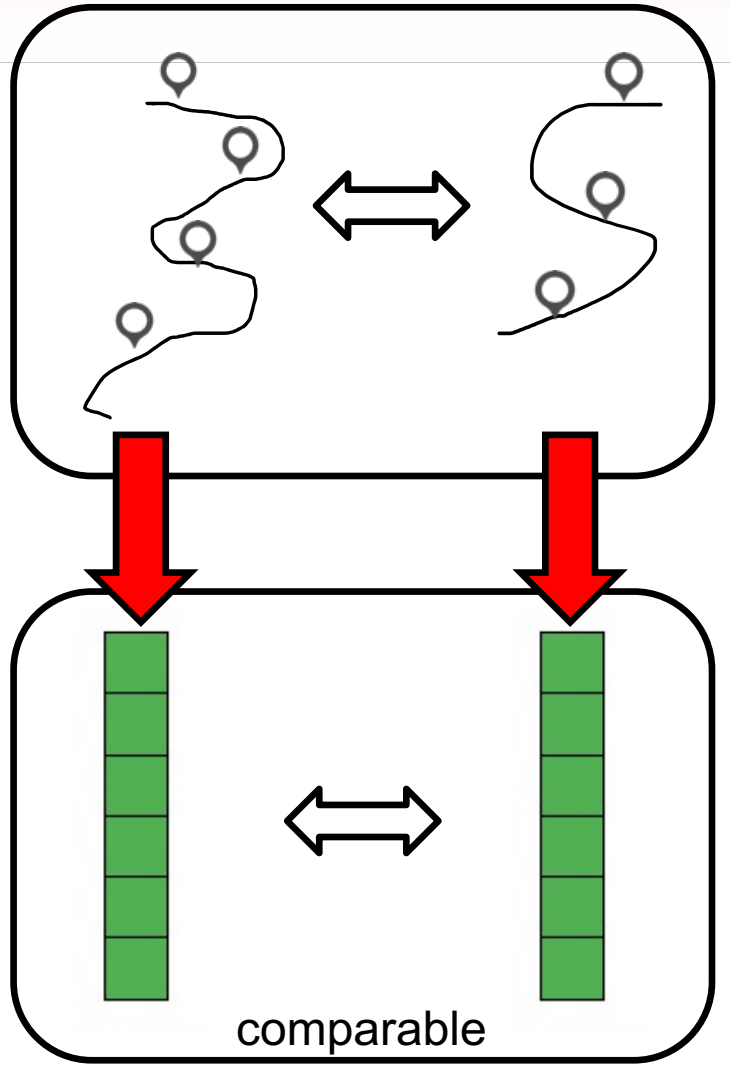
NAIST

Human Movement Trajectories

- The ubiquity of such wearable devices as smartphones continues to deepen its presence in modern societies
- Possible to analyze and visualize people who are moving as part of a trajectory of GPS big data
- Cluster human movement trajectories using time-series distributional representations
- Calculated the distance of the representation vectors derived from neural network models

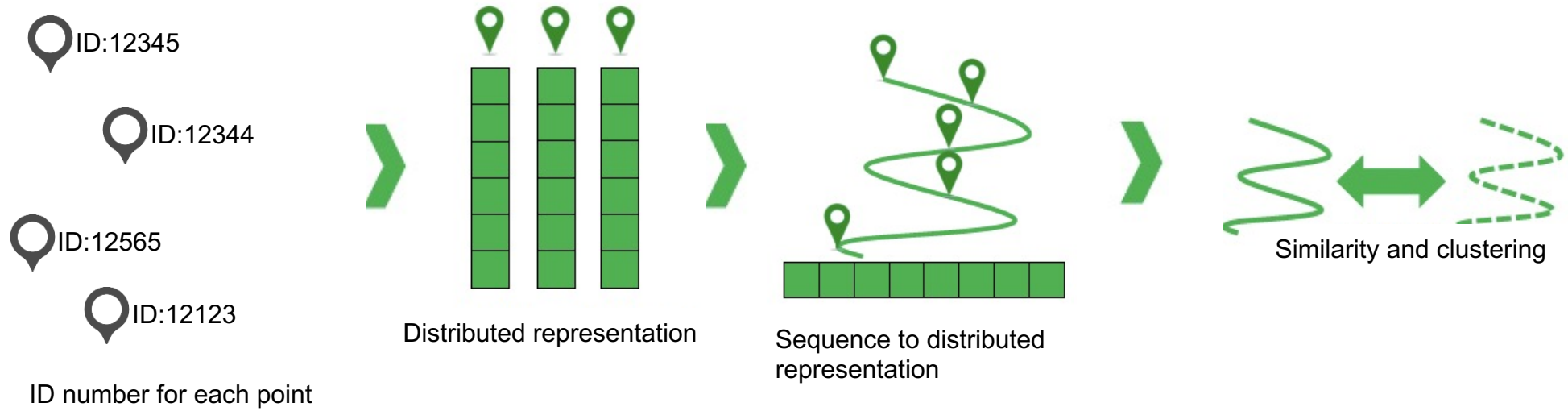
Distributional representations

- By analyzing the trend of visiting routes, it will be utilized for effective tourism strategies and product development
  - Problem: The amount of data is huge and diverse, making it difficult to compare and analyze
- The length of the pathways and the places they pass through are varied and differed
  - Fixed-length vectorization of pathway: distributional representation



Clustering based on distributional representations

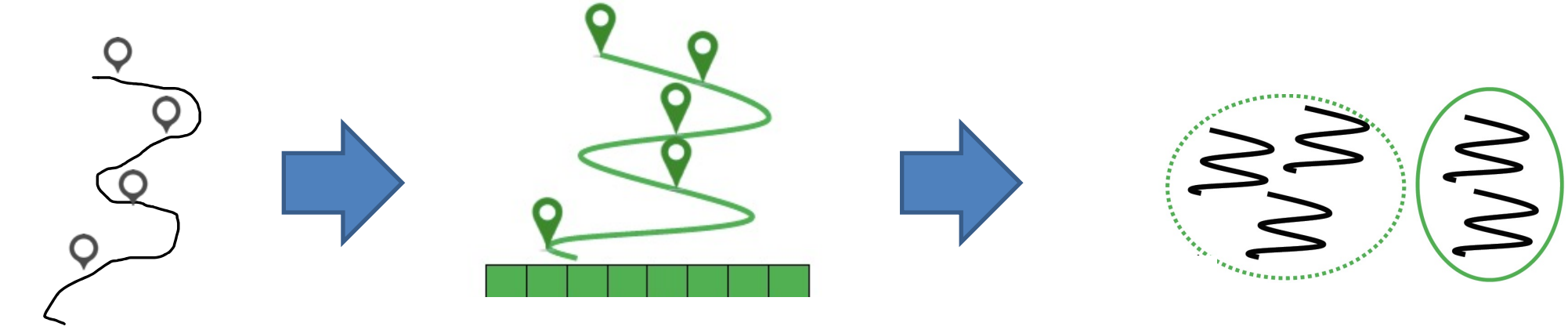
- Human movement analysis by similarity of time series using distributed representation [Crivellari+ 2019]
  - Consider behavioral proximity (semantic closeness of moving sequences)
  - Based on Mesh2Vec (Word2Vec applied to a human movement trajectories)



Applying bi-directional LSTM for clustering

Baseline of this study

- Human moving prediction using long short-term memory (LSTM) [Crivellari+ 2019]
- Obtaining distributed representation of human moving series by using Bi-LSTM [Kubo et al., 2020]
  - It does not distinguish between forward and backward directions because our purpose is not only to predict but to analyze human behaviors
  - Grouping similar distributed representations of time series by using hierarchical clustering



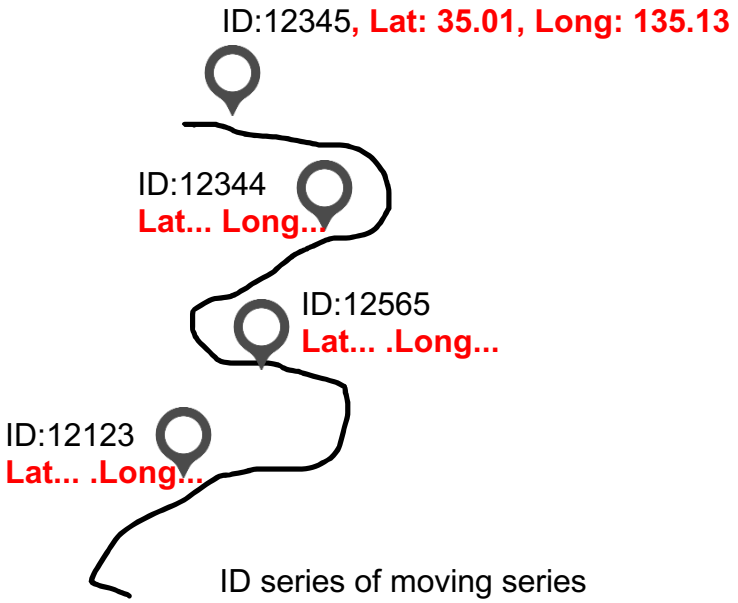
ID series for each location

Clustering users based on similarity

Objective of this research

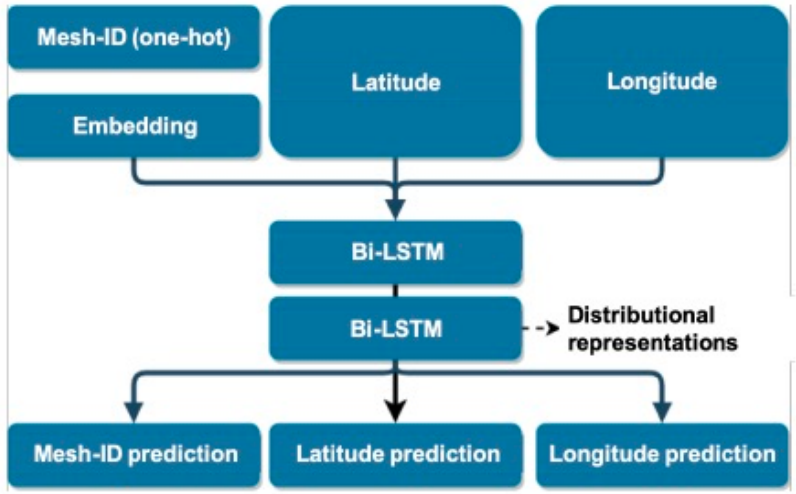
- Previous work
  - Use only discrete mesh-IDs as an input
- Problem in previous research
  - The geographical coordinates (latitude and longitude) of each point is not explicitly shown

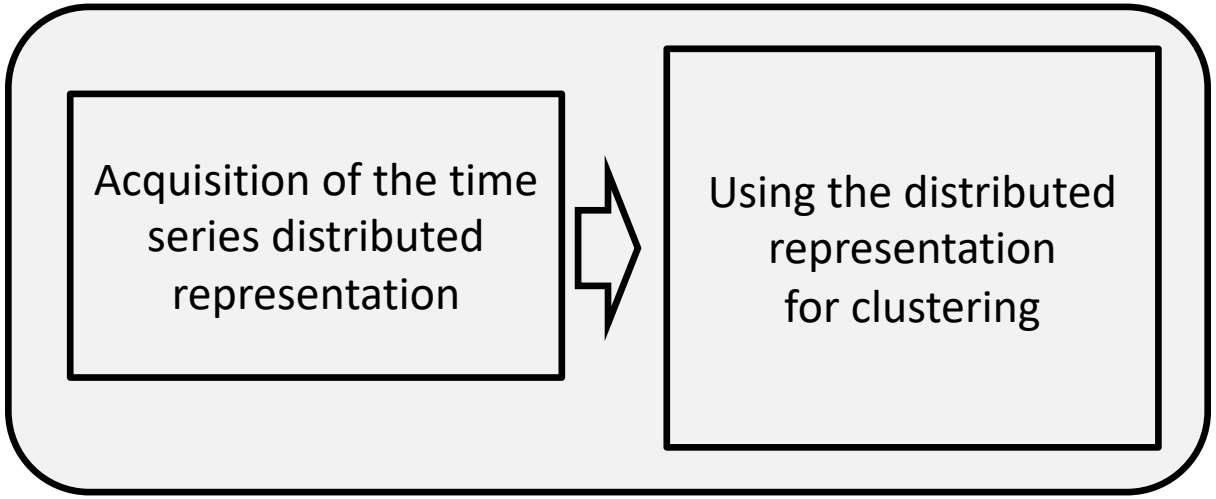
Acquisition of distributed representation based on discrete mesh-ID and continuous latitude/longitude Information and its validity through hierarchical clustering



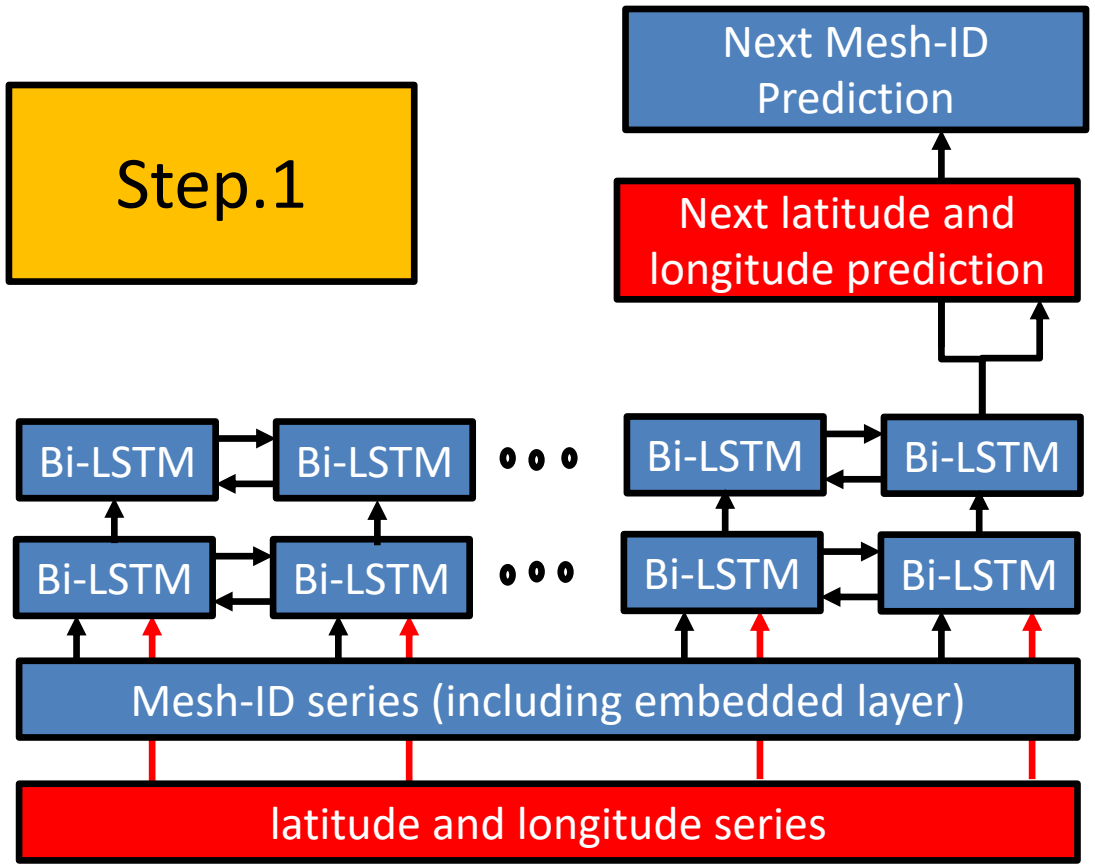
Proposed approach

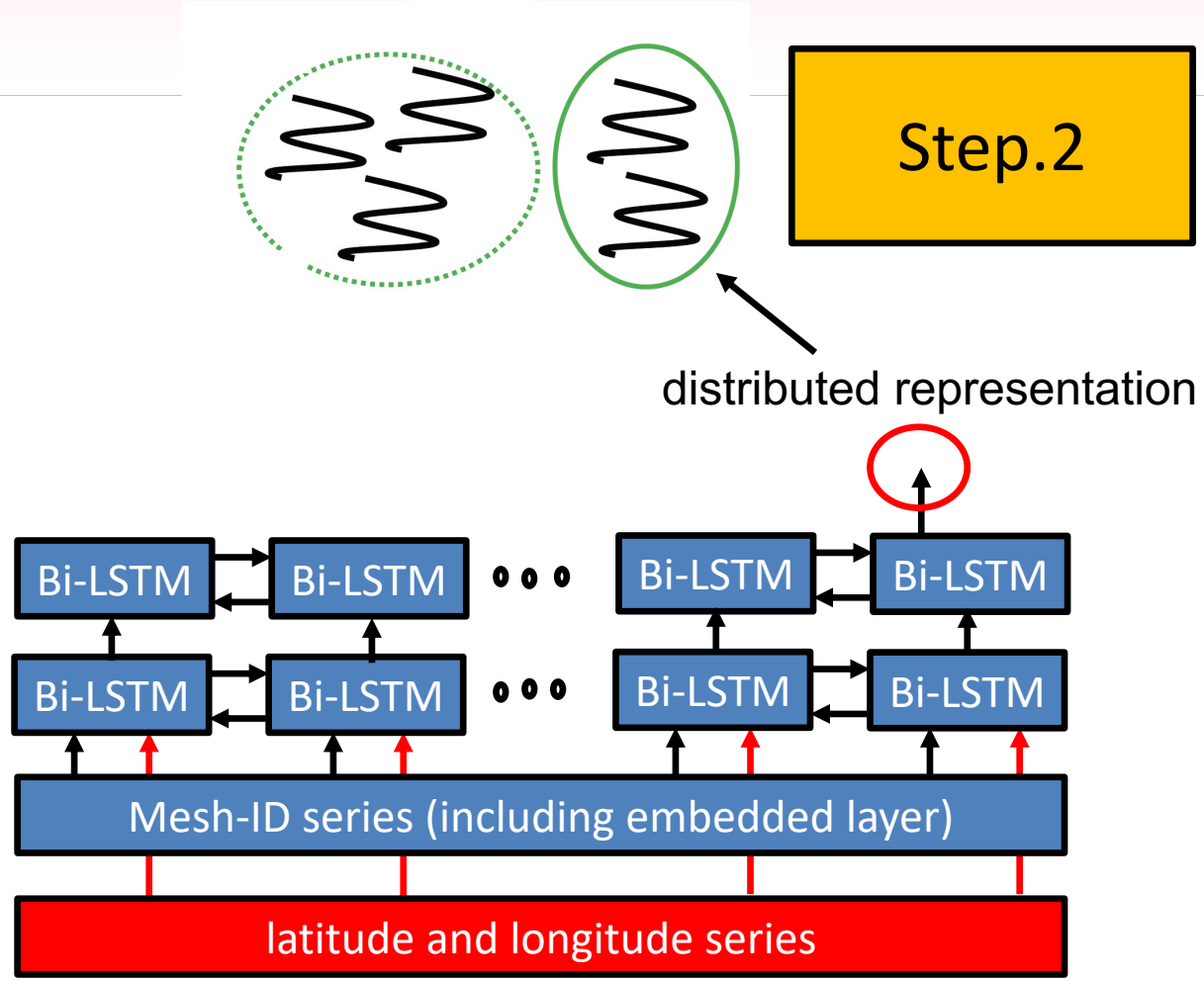
- Objective: Human movement clustering based on distributed representations
- Proposal
  - Acquisition of distributed representation of time series by Bi-LSTM with additional geographical coordinates as the input
  - Effect of adding geographical coordinates using hierarchical clustering



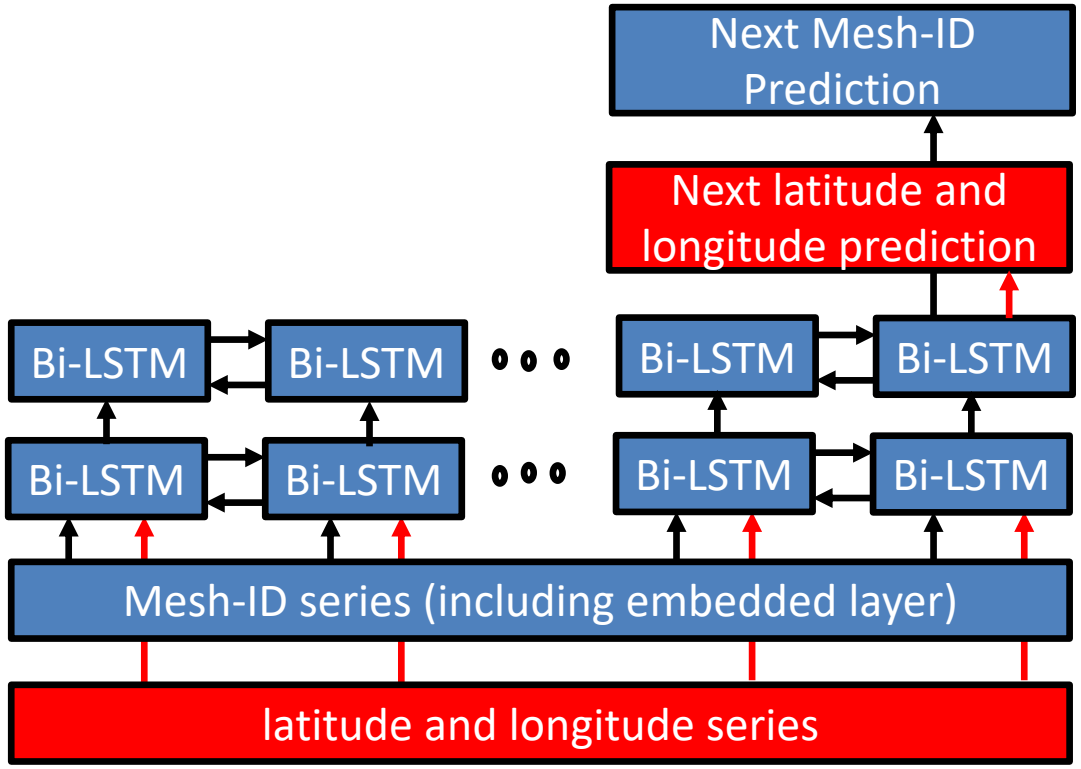


Predication and clustering





Details of the prediction model



Prediction Target :
- Last Mesh-ID
- Last Latitude (Lat)
- Last Longitude(Long)

Loss function

$$L_{total} = \alpha * L_{mesh} + \beta * (L_{lat} + L_{long})$$

L_{mesh} : Cross entropy loss
$L_{\{Lat,Long\}}$: Mean squared error
α, β : constants (1,1)

Experimental evaluation

- Dataset (provided by Agoop Corp)
  - Target areas: Tokyo, Japan
  - Mesh size : 100m squared
  - Total number of meshes (calculated from the area): 219,396
  - Period: January to March 2021
  - Data separation ratio for training, validation, and testing: 8:1:1
  - Number of users for training : 35,888
  - Number of users for clustering : 36 (only more than 200 series length)

- Hierarchical clustering
  - Cluster determination method: Ward's method
  - Distance measure: cosine distance
  - Cluster distance threshold: 1.2

Accuracy – last mesh prediction

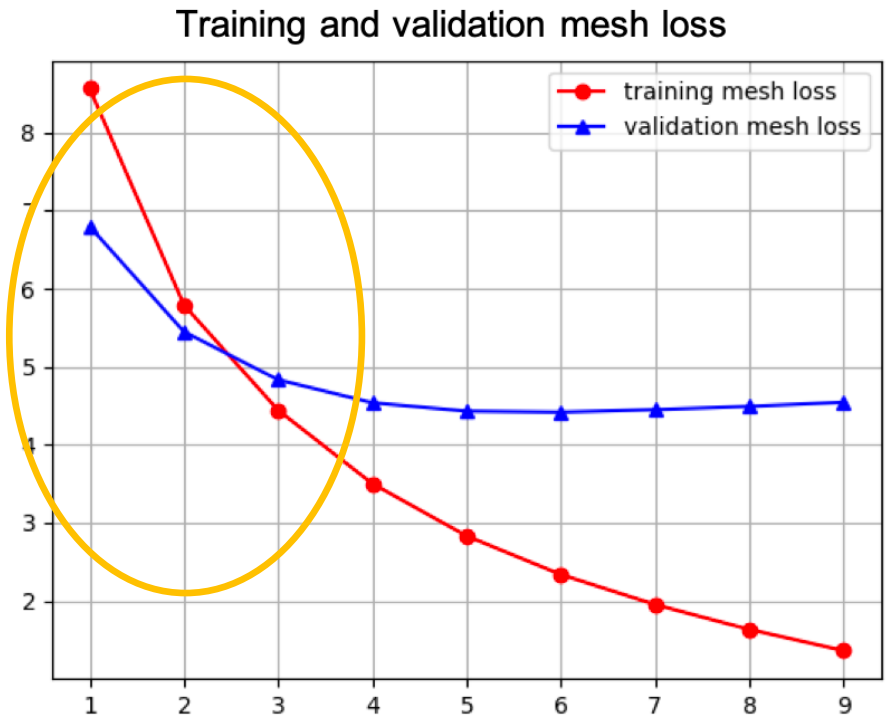
We obtained more than 30% of accuracy (baseline: 21.5%)

- Because the total number of mesh-IDs in Tokyo prefecture is 21,939 the random prediction would obtain about 0.004% of accuracy
- We had another experiment in data of the Kyoto area where the total number of mesh-IDs is around 8000
- In such data, we obtained more than 50% of accuracy

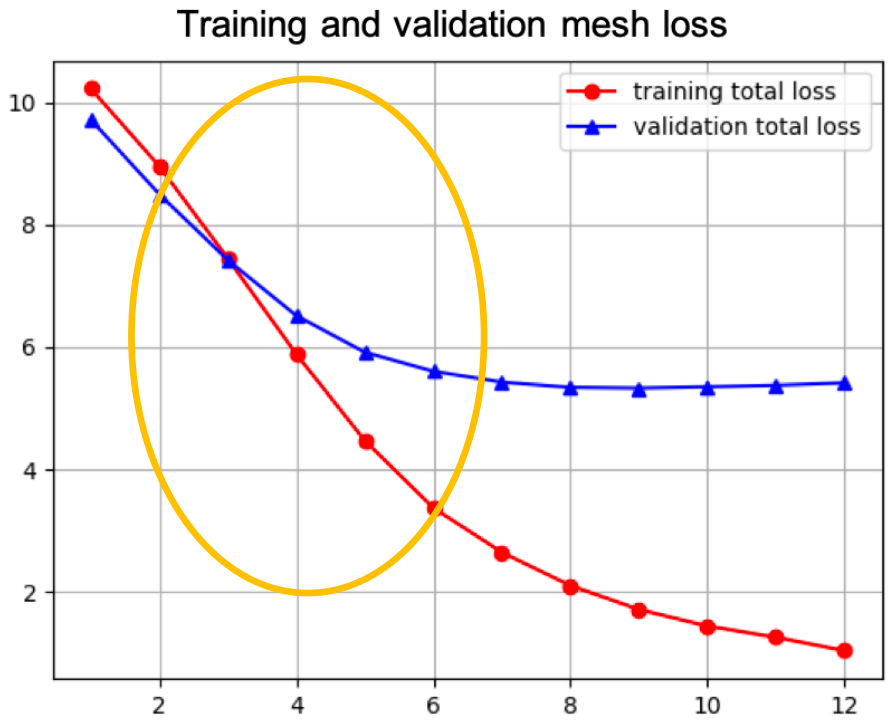
Model	Accuracy %
Mesh-ID-only (LSTM network)	21.5
Mesh-ID-only (Bi-LSTM network)	29.0
Mesh-ID plus latitude/longitude (LSTM network)	33.8
Mesh-ID plus latitude/longitude (Bi-LSTM network)	30.8

Loss curve

The addition of latitude/longitude reduced the loss faster in the early stage



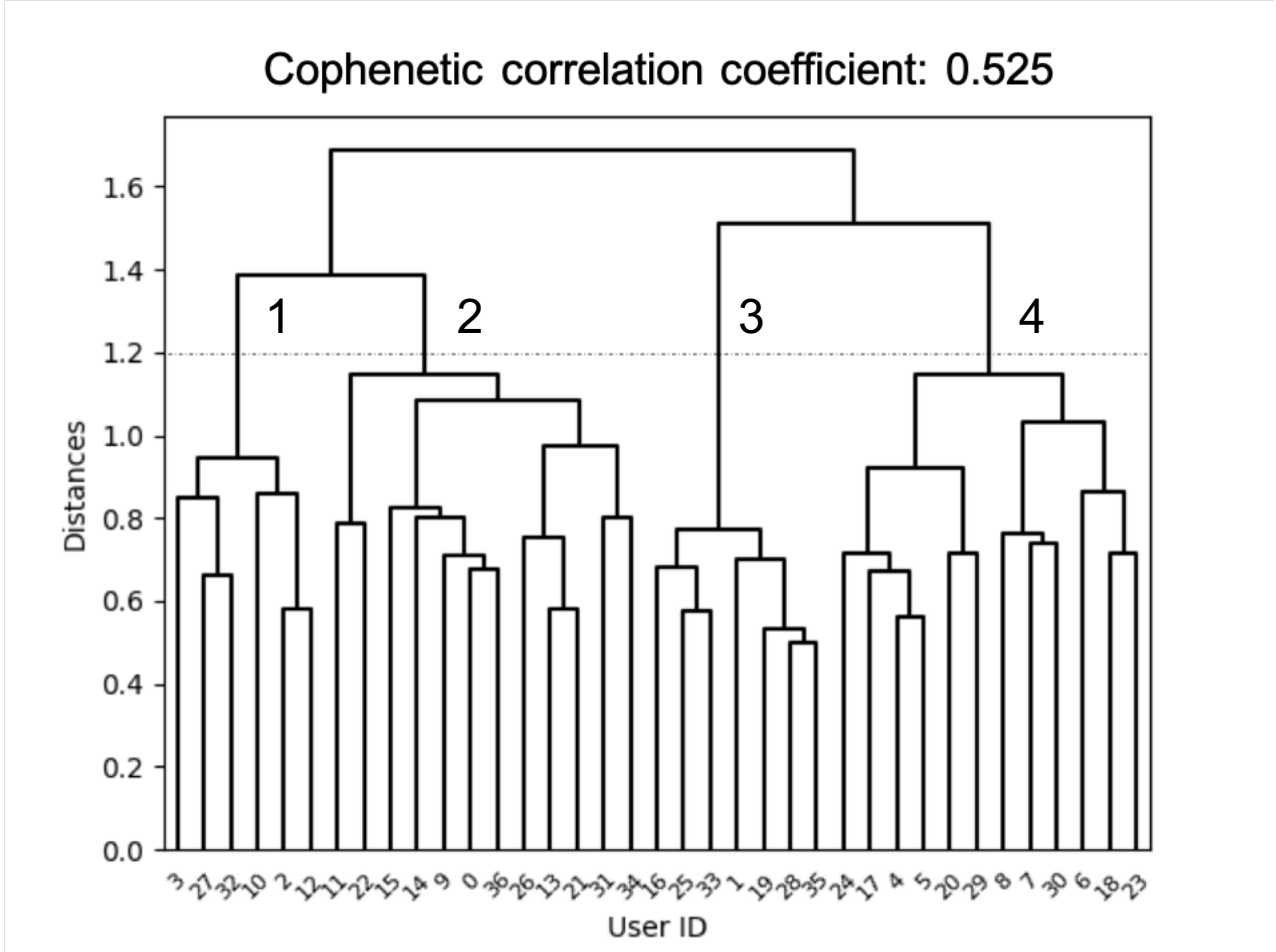
Mesh-ID cross entropy loss in a model of Mesh-ID plus latitude/longitude



Mesh-ID cross entropy loss in a model of Mesh-ID-only

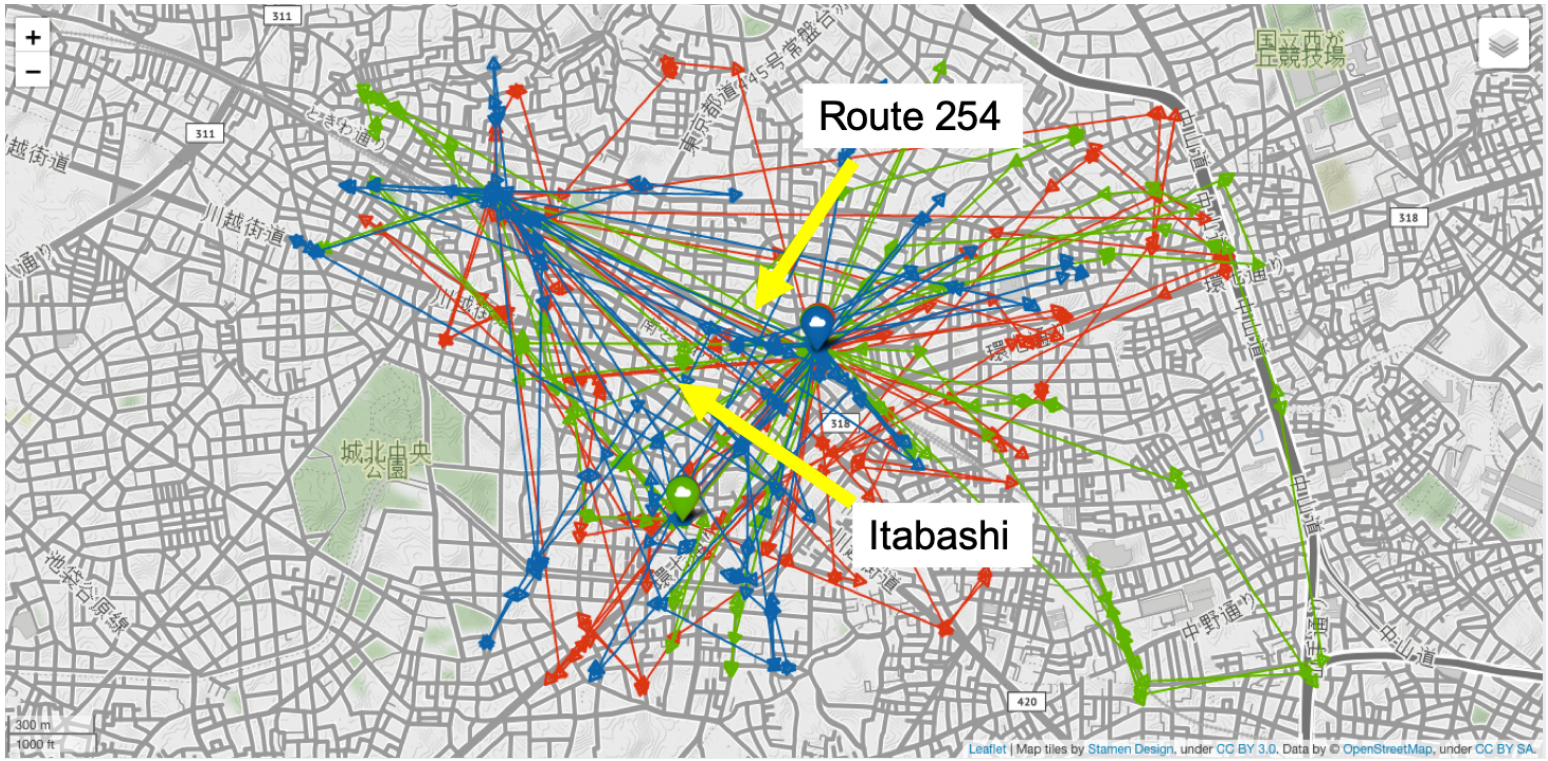
Clustering

Clustering in a model of Mesh-ID plus latitude/longitude: four clusters



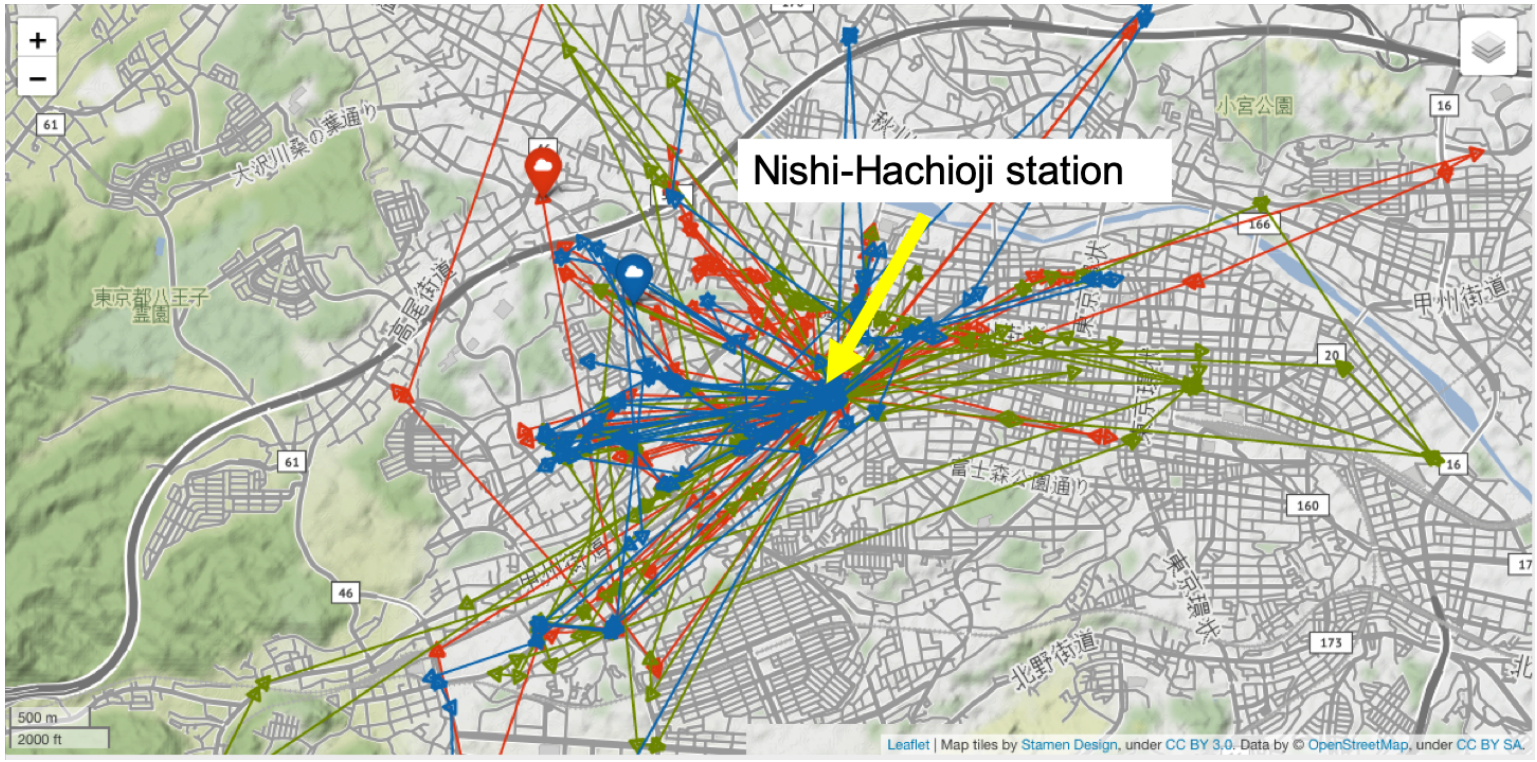
Cluster visualization (1)

Clusters number 2 with top three nearest
- These users moved around Itabashi-Ku and route 254



Cluster visualization (2)

Clusters number 3 with top three nearest
- These users moved around Hachioji and Nishi-Hachioji station



Conclusions

- Proposed Bi-LSTM network and integrated additional geographical coordinates (latitude and longitude information) into models to accurately predict the last mesh and constructed clusters based on distributional representations
- Future work
  - The current method must conduct a heuristic analysis and researchers need to observe it, which is time-consuming and expensive
  - To solve this problem, an objective analysis must be studied that incorporates such landmarks as points of interest (e.g., popular stores) and geographical features that are likely to be useful for predicting human movement tendencies