Clustering post-hurricane human mobility using spatial graphs and LLM-derived semantic embeddings

Sizhe Zhang [1] Xinyu Zhang [2] Kenan Li [3] Nan Lin [1]

[1]Washington University in St. Louis [2]Boston University [3]Saint Louis University

Introduction and Research Data

Understanding human mobility after major meteorological events is critical for disaster response. We analyzed cellphone mobility data from the 143 hours following Hurricane Ian's landfall using a novel unsupervised framework. We selected the mobility paths that originated in the landfall zone of Hurricane Ian. To standardize each path, the recorded GPS locations within every given hour were averaged into a single representative locations. This research introduces a novel unsupervised framework to cluster sparse mobility trajectories by integrating spatial network context with semantic Point of Interest (POI) information.

Methodology

We tessellate the study area into hexagons and derive two 64-dim embeddings per cell via **Node2Vec** [Grover and Leskovec(2016)] and nonlinearly dimension-reduced **Gemma3 LLM** embedding [Team(2025)].

$$h : \mathbb{R}^2 \to \{1, \dots, H\}, \quad G = (V = \{1, \dots, H\}, E)$$

$$g_i = \mathrm{Node2Vec}(G)_i, \quad s_i = \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \mathrm{AE}(\mathrm{Gemma3}(p)), \quad x_{n,t} = \left[s_{h(p_{n,t})} \,\|\, g_{h(p_{n,t})}\right] \in \mathbb{R}^{64+64}$$

These vectors are modeled by a **Attention-based Transformer Autoencoder**, with loss at each epoch given below:

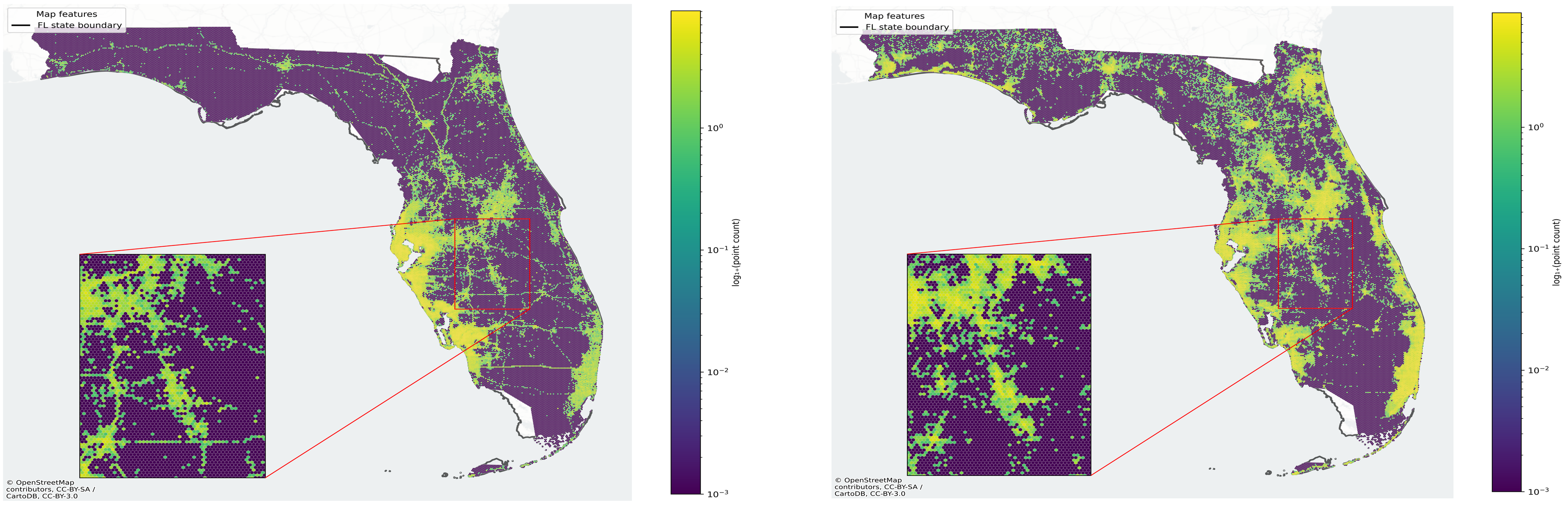
$$z_n = f_{\mathrm{enc}}(x_{n,1:T}), \quad \hat{x}_{n,t} = \left[f_{\mathrm{dec}}(z_n)\right]_t$$

$$\mathcal{L}^{(e)} = \underbrace{\frac{1}{\sum_{n,t} m_{n,t}} \sum_{n,t} m_{n,t} \|x_{n,t} - \hat{x}_{n,t}\|_2^2}_{\text{masked MSE}} + \underbrace{\lambda_0 \min\left(1, \frac{e}{E_{\mathrm{warmup}}}\right) \frac{1}{N} \sum_n \|z_n\|_2^2}_{\text{warm-up } L_2}$$

The final step is a k-Means clustering on the latent space of the Machine Learning model above.

$$\{z_n\} \xrightarrow{\text{cluster}} \{c_n\}, \quad c_n = \arg\min_k \|z_n - \mu_k\|^2$$

Fig. 1 shows a strong spatial correlation between trajectory points (a) and Points of Interest (b). This overlap allows us to use POI data as a proxy to infer travel intent [Nadiri *et al.* (2025)]. We also incorporate spatial graph embeddings to provide essential connectivity context for our analysis [Guan and Chen(2021)].



(a) Trajectory point counts on a hex grid. (b) Point-of-interest counts on the same grid.

Figure 1. Spatial distributions of POI and trajectory points across Florida, aggregated on a 1.25km hexagonal grid. The inset zooms into [27°39'N, 81°30'W] to highlight local density patterns.

Framework Overview

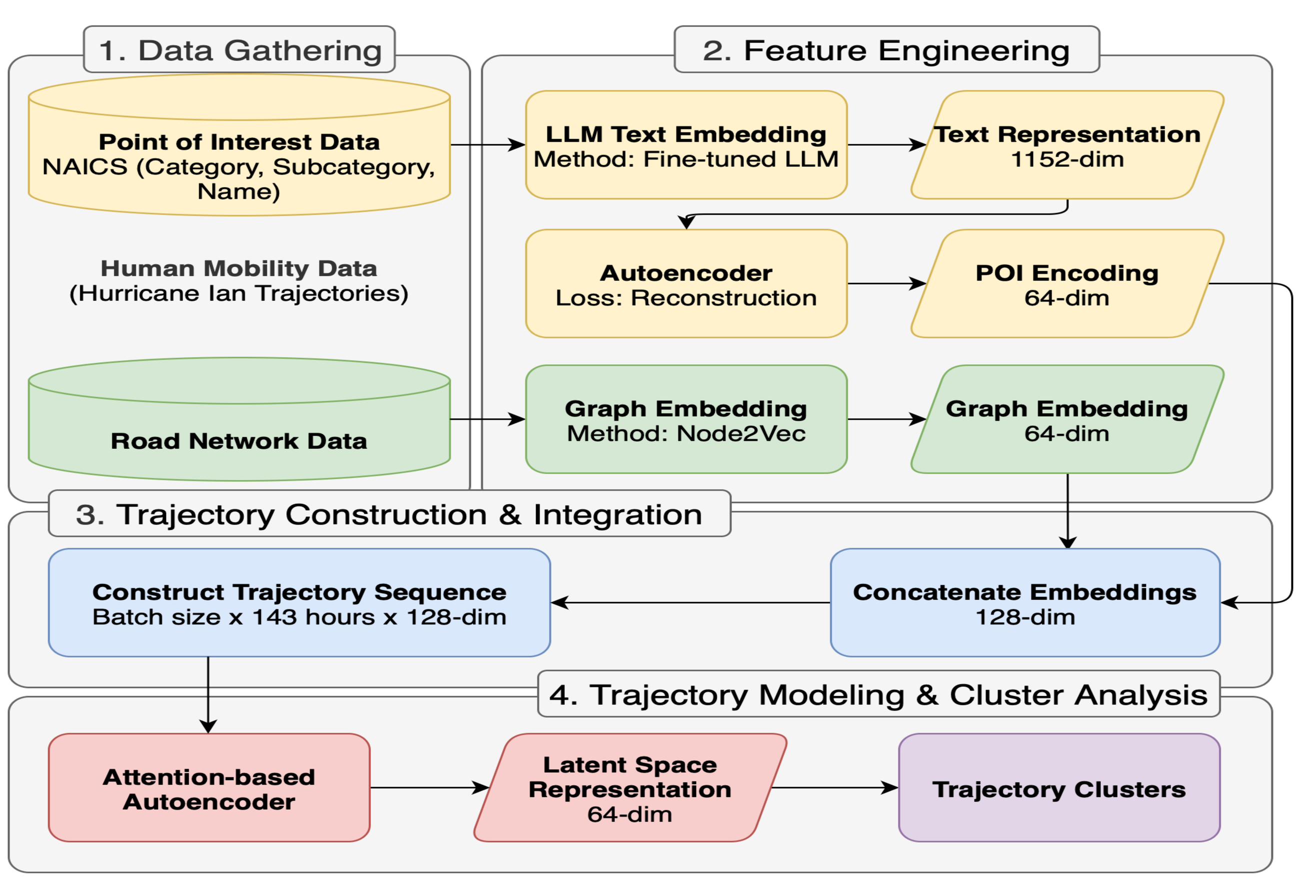


Figure 2. Unsupervised Trajectory Clustering Pipeline

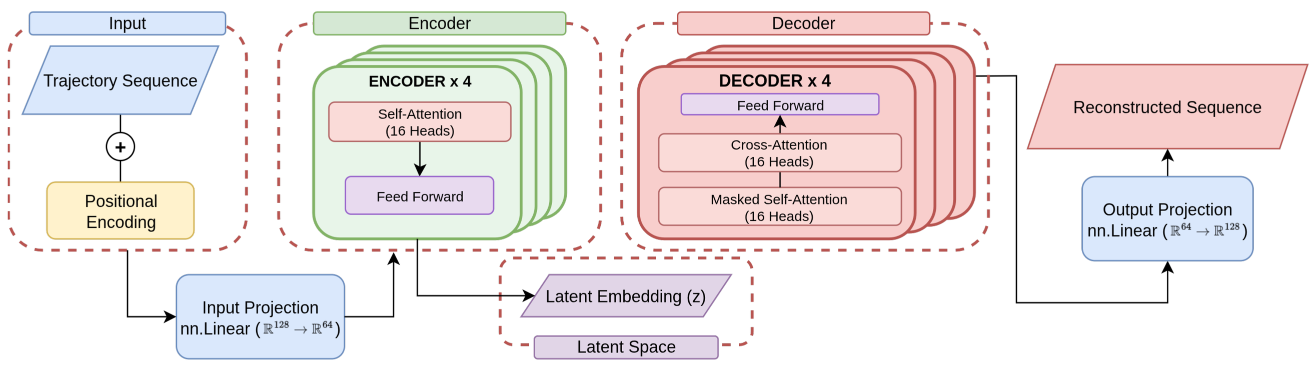
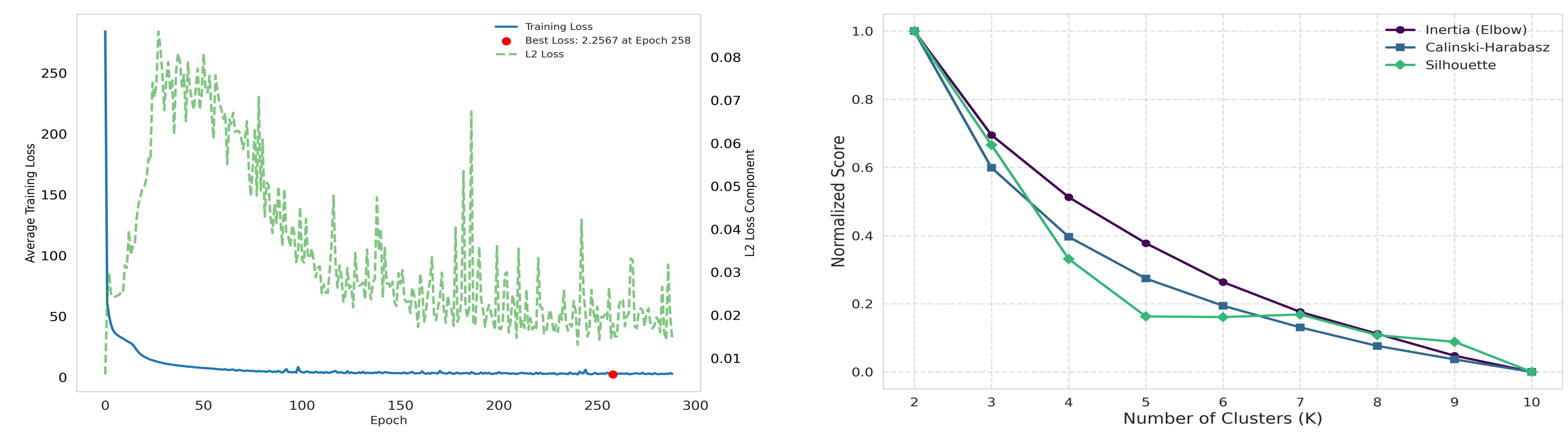


Figure 3. Trajectory Autoencoder Model. LayerNorm blocks are omitted for simplicity.

Results

Fig. 4a shows the change of L_2 regularization and total training loss over training. The model exhibits rapid initial convergence. The best model, marked in red, was saved at epoch 258, and training was halted at epoch 288 after 30 epochs without improvement. L_2 regularization loss also decreases steadily over training. We then proceed to perform clustering on the latent space $\{\vec{z}_n\}$.



(a) Training loss curve for the Trajectory Transformer Autoencoder. The plot shows the combined training loss (blue curve) and the L2 regularization loss (green dashed line) for each epoch.

(b) Metrics for choosing the number of clusters. The different criterions are normalized and converted such that a higher score indicates a more optimal cluster number.

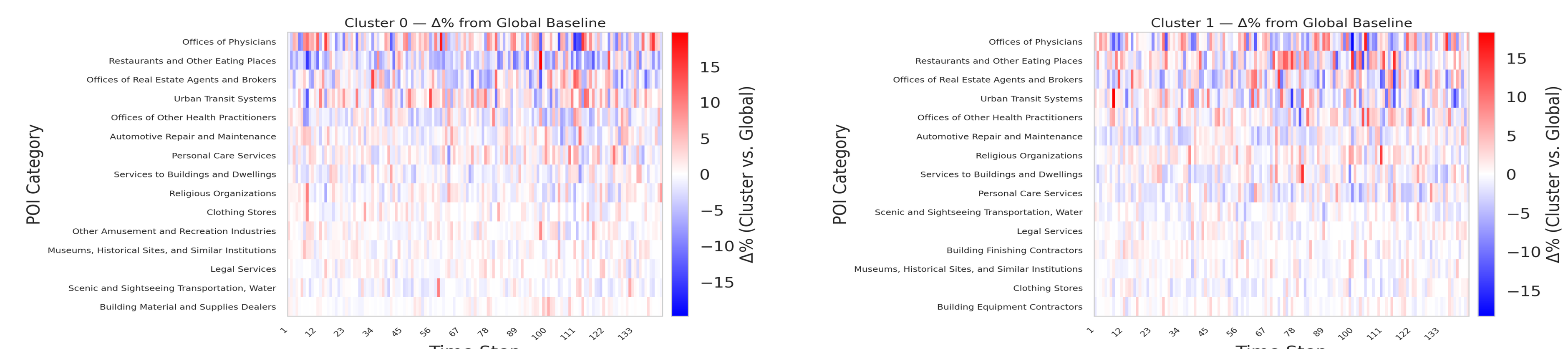
Figure 4. Training results and clustering metrics

Fig. 5a, 5b, and 5c visualize the unique temporal signature of a trajectory cluster. For trajectory n in cluster i, the $\Delta\%$ value associated with category p at time t is given as:

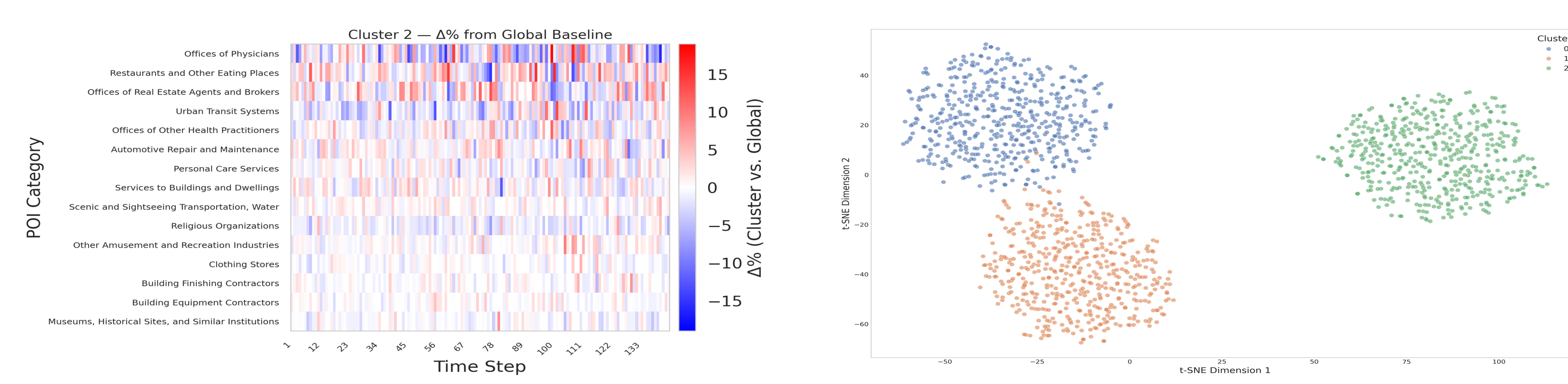
$$P_{\mathrm{cluster}}(n, p, t) = \frac{C_{n,p,t}}{\sum_{p' \in P_{\mathrm{Top15}}} C_{n,p',t}}, P_{\mathrm{global}}(p, t) = \frac{\sum_m C_{m,p,t}}{\sum_{p' \in P_{\mathrm{Top15}}} \sum_m C_{m,p',t}}$$

$$\Delta\%_{i,p,t} = (P_{\mathrm{cluster}}(i, p, t) - P_{\mathrm{global}}(p, t)) \times 100$$

The $\Delta\%$ allows us to define each cluster by its most significant activities over time.



(a) POI occurrence heatmap for Cluster 0. (b) POI procurance heatmap for Cluster 1.



(c) POI occurance heatmap for Cluster 2. (d) t-SNE plot for the 3 clusters. The top $M = 500$ representative trajectories were selected from each cluster.

Figure 5. Training results and clustering metrics

Conclusions

1. **Robust Unsupervised Clustering:** We employ a fully unsupervised framework to uncover statistically distinct clusters directly from raw mobility trajectories. The resulting latent embeddings exhibit high inter-cluster separation, demonstrating the model's capacity to dissect complex, unlabeled mobility patterns.
2. **Enhanced Interpretability & Causal Modeling Potential:** By projecting trajectories into a concise latent space, our method yields interpretable variables that can be readily incorporated into downstream statistical analyses. These latent components form a rigorous basis for future causal-inference studies.

References

[Grover and Leskovec(2016)] Aditya Grover and Jure Leskovec. node2vec: Scalable feature learning for networks, 2016. URL https://arxiv.org/abs/1607.00653.

[Guan and Chen(2021)] Xiangyang Guan and Cynthia Chen. A behaviorally-integrated individual-level state-transition model that can predict rapid changes in evacuation demand days earlier. *Transportation Research Part E: Logistics and Transportation Review*, 152: 102381, 2021. ISSN 1366-5545. doi: https://doi.org/10.1016/j.tre.2021.102381.

[Nadiri et al.(2025)Nadiri, Li, Faraji, Abuoda, and Papagelis] Amirhossein Nadiri, Jing Li, Ali Faraji, Ghadeer Abuoda, and Manos Papagelis. Trajlearn: Trajectory prediction learning using deep generative models. *ACM Trans. Spatial Algorithms Syst.*, 11(3), May 2025. ISSN 2374-0353. doi: 10.1145/3729226.

[Team(2025)] Gemma Team. Gemma 3. 2025. URL https://goo.gle/Gemma3Report.

Funding Information

This research is supported in part by NSF DMS-2418979 and a Washington University Geospatial Seed Grant.
Placeholder1

sizhe@wustl.edu