Trajectory Test-Train Overlap in Next-Location Prediction Datasets

Keywords: Human Mobility, Next-Location Prediction, Generalization, Test-Train Overlap

Next-location prediction is the task of forecasting which location an individual will visit, given their historical trajectories. It is crucial in many applications such as travel recommendation, and optimization, early warning of potential public emergencies, geomarketing, and others. Deep learning-based next-location predictors (NLs) and has driven test-set performance on mobility data to new heights [1]. However, little work has been done on how challenging these benchmarks are, what NLs learn, and their actual generalization capabilities. Although some studies investigate the predictability of human whereabouts and its relationship with the trajectories' features [2], we know comparatively little about how the individuals' trajectories are distributed in mobility benchmarks, making it hard to understand and contextualize our observed results. Recent studies in natural language processing and computer vision show that deep learning models excel on specific test sets but are not solving the underlying task.

In this paper, we investigate whether it is the case for NLs too. To address this compelling issue, we stratify mobility data by whether the trajectories in the test set also appear fully or partially in the training set. We explore three ways to quantify overlap: Jaccard Similarity (JS), Longest Common Subsequence (LCST), and Overlap From the End (OFE). Jaccard Similarity (JS) measures the percentage of locations in the test trajectories that are also in the training trajectories, regardless of the order in which locations appear. The Longest Common SubTrajectory (LCST) is the longest subtrajectory in common between two trajectories. The Overlap From End (OFE) enforces that the common subtrajectory is at the end of the two trajectories.

In Figure 1.a, we can see the percentage of test trajectories overlapping with training trajectories for a given percentage on two specific datasets. Experiments on the other datasets can be find in [3] In general, we find that, in five next-location benchmark datasets, there is a severe problem of trajectory overlapping between the test and training sets when composing them randomly: $\sim 43\%$ to 72% of test trajectories overlap at least with 50% of the points with trajectories in the training set, and with 7% to 14% of test sub-trajectories entirely overlap training sub-trajectories. In other words, based on the standard way training and test sets are split in the literature, a significant portion of the trajectories in the test sets have already been seen during training.

Based on these observations, we propose to evaluate NLs on *stratified test sets based on the overlap between trajectories in the training set*. The models are evaluated in terms of accuracy-at-5 (ACC@5) i.e., whether or not the actual next location is within the 5 locations with higher probabilities indicated by an NL. Figure 1.b shows significant variability in model performance, varying the percentage of overlap. Indeed, we find an accuracy $\leq 5\%$ when predicting unseen trajectories (novel mobility) and $\geq 90\%$ when predicting trajectories with high overlaps (known mobility). Surprisingly, we also find that DL-based NLs perform even worse than baseline models (Markov Chains) when tested on novel mobility. Our results are consistent across the datasets analyzed and the NLs selected, demonstrating that current train/test splits are flawed, and more robust methods are needed to evaluate the generalization capabilities of NLs.

Also, a possible reason why NLs perform poorly on trajectories is that RNNs focus on memorizing regularities in long sequences, thus limiting NLs' generalization capabilities. Wrong location predictions may happen, for example, when the NL's probabilities assigned to each potential location (i.e., the locations' scores) are relatively uniformly distributed. Entropy, by definition, can be used to measure how much distribution is uniformly distributed. For instance, if the softmax is supposed to output a probability for N classes, having a score of 1 means that all the classes have an equal probability of $1/N$ to be chosen as the next location. Our findings in **??** suggest that NLs may have to deal with uncertain and uniformly distributed outputs of the softmax layer when dealing with novel mobility traces. Our intuition is that while a location in the test set may not (or rarely) appear in the training set, we may exploit well-known mobility laws to capture an individual's behavior and predict the location they will potentially visit and reduce the entropy score through re-ranking techniques that account for mobility laws. The results are showed in Figure 1.c with the original accuracies in blue and the new ones in red after the injection of three mobility laws in the reranker: distance law, visitation frequency law, and returner/explorer dichotomy. The accuracies are related to the to the trajectories with 0-20% overlap. This work finds that the models' performances are deeply affected by the level of overlap in the test trajectories. Based on the amount of trajectory overlap, we identify three scenarios: (i) **Known Mobility** (high overlap levels): predictive performance is much higher than the performance on a non-stratified test set (close to 100%) as the test trajectories are almost identical to the training trajectories. (ii) **Fragmentary Mobility** (overlap between 20% and 80%): the majority of trajectories in the test set lie in this scenario. There is a drop in the model performance compared to the previous scenario, decreasing to $\sim$80%. (iii) **Novel Mobility**(overlap below 20%): A significant number of trajectories lie in this scenario. However, since NLs cannot rely on the trajectories already seen in the training phase, these are the most difficult trajectories to predict. Indeed, the performance of NLs on test sets with low overlap is considerably lower than on a non-stratified test set. While predicting known mobility is a simple task, inferring mobility patterns for fragmentary mobility and novel mobility presents challenges (e.g., dealing with under-represented locations or not represented at all in the training set). From a modeling perspective, this may suggest that current models are excellent in memorizing already seen trajectories but cannot generalize well. Some works suggest that reranking techniques or few-short learning algorithms may help solve this problem. Also, results indicated that NLs might not be evaluated adequately. In this sense, here we provide a set of recommendations for the evaluation of NLs: (i) MMCs achieve performance similar to NLs. Therefore, we claim that MMCs and other Markov chain approaches should always be used as a baseline. (ii) Although NLs achieve good overall performance, they are biased due to trajectory overlap. Besides the NLs' average performance, researchers should report the performance for the known mobility and the novel mobility scenarios. It is indeed crucial to understand whether the improved performance of the proposed NL is actually due to its increasing generalization capability or because it is memorizing better the trajectories in the training set; (iii) NLs achieve the worst performance on the 0-20% overlap bin. We can improve the performance on this bin, hence increasing NLs' generalization capability with the support of well-known spatial mobility laws, which are loosely captured by state-of-the-art NLs given their reliance on RNNs.

References

[1] Massimiliano Luca, Gianni Barlacchi, Bruno Lepri, and Luca Pappalardo. A survey on deep learning for human mobility. *ACM Computing Surveys (CSUR)*, 55(1):1–44, 2021.

[2] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. Limits of predictability in human mobility. *Science*, pages 1018–1021, 2010.

[3] Massimiliano Luca, Luca Pappalardo, Bruno Lepri, and Gianni Barlacchi. Trajectory test-train overlap in next-location prediction datasets. *arXiv preprint arXiv:2203.03208*, 2022.

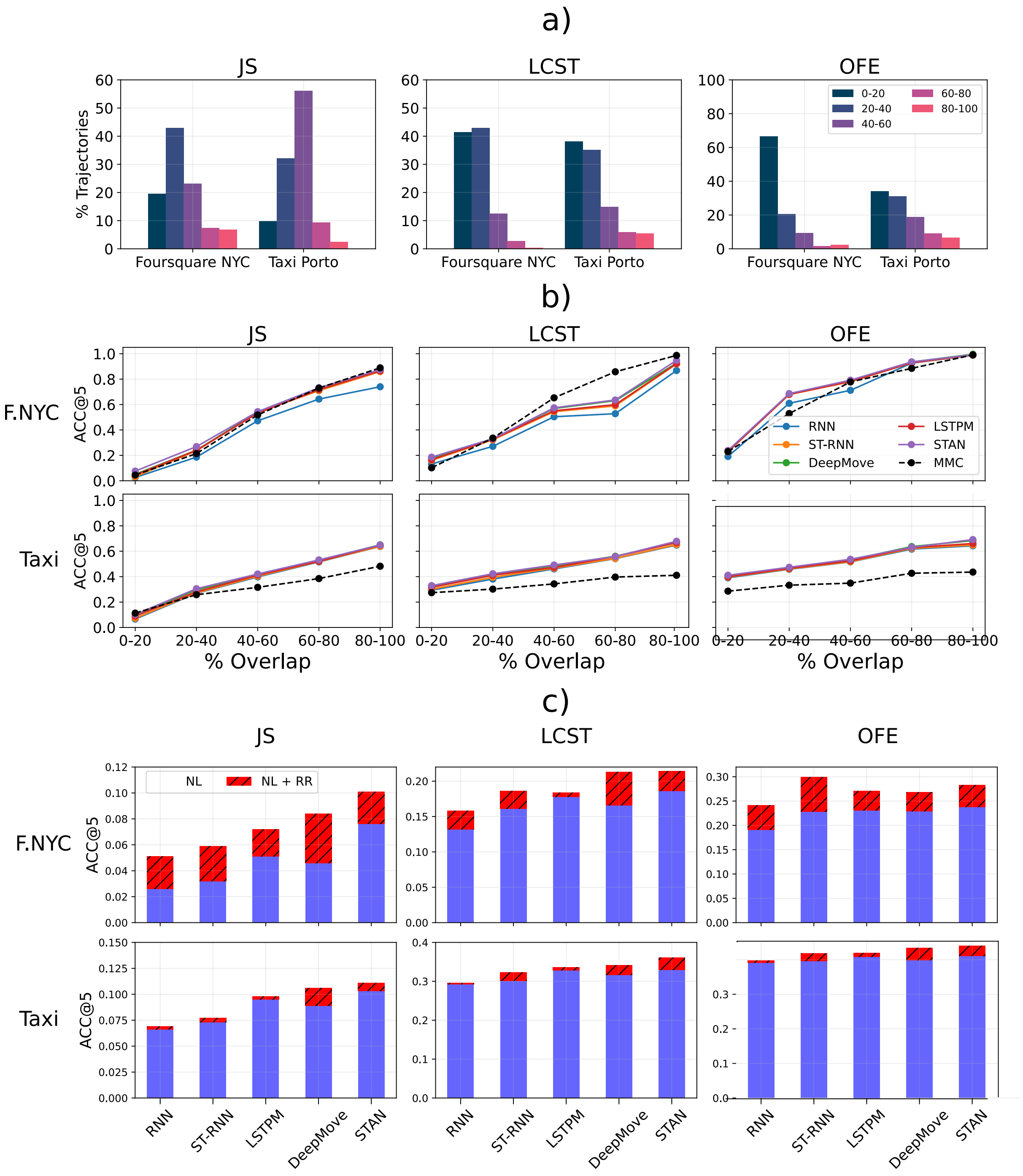

Figure 1: a) the percentage of test trajectories overlapping with training trajectories for a given percentage. b) The ACC@5 of all the models taken into account when tested only on trajectories with a certain range of overlap. c) the original ACC@5 in blue and the ACC@5 after a reranking with mobility laws in red. We report only the ACC@5 of trajectories with 0-20% overlap