Extracting Visited Points of Interest from Vehicle Trajectories

Ilkcan Keles
Simonas Saltenis
Christian S. Jensen
Aalborg University

Matthias Schubert
Peer Kröger
Ludwig-Maximilians Universität München

Outline

- Introduction
- Problem Definition
- Related Work
- Visited PoI Extraction Method
  - Overview
  - Building the Bayesian Network
  - Assignment
- Experimental Evaluation
- Conclusion

Introduction

- More and more GPS data is collected from vehicles.
- Users' visits to PoIs can be extracted from this data.
- These visits offer insights into the PoIs.
  - Popularity
  - Importance
  - The duration of visits
  - This information can be extracted at different spatial and temporal granularities.
    - The popularity of a PoI to visitors coming from a specific region.
    - The popularity of a PoI to visitors in the morning.

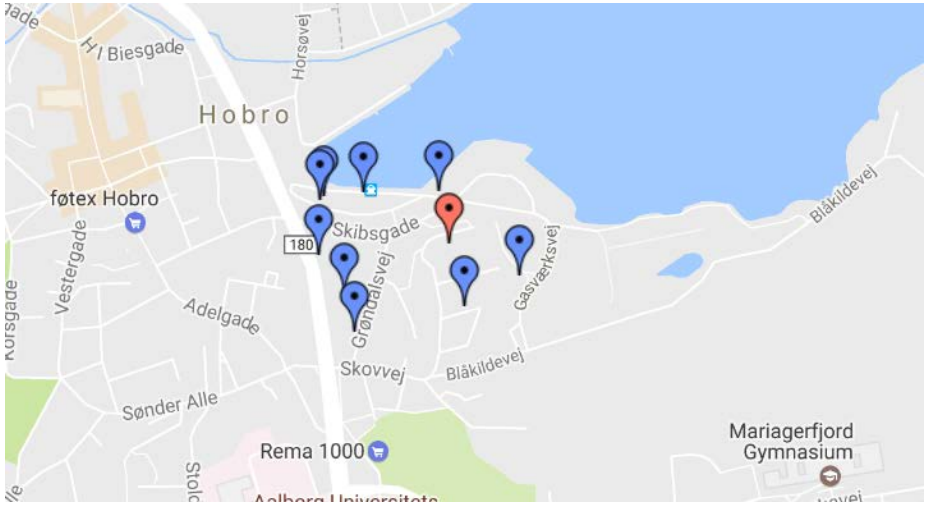
Problem Definition

- S_{TR} – A set of GPS trajectories
- S_P – A database of PoIs in the geographical region covered by S_{TR}
- Given S_{TR} and S_P, the problem is to identify the visits of users whose trajectories are given in S_{TR} to the PoIs contained in S_P.
  - Two subproblems: Identifying the stops in the trajectories and assigning the stops to PoIs
  - We focus on the second subproblem.

Related Work

- Enriching trajectories with semantic information
  - SMoT and CB-SMoT annotate stops.
  - Many proposals use clustering to identify interesting and significant locations.
  - Battacharya et al. propose a method based on bearing change, speed and acceleration to identify interesting places.
- Extracting visited PoIs and activities from GPS trajectories
  - Nishida et al. propose a probabilistic PoI identification method .
    - Semi-supervised
    - A hierarchical Bayesian model that makes use of personal preferences, stay locations, and stay times for each PoI category
  - Bhattacharya et al. propose a two-phase algorithm for assignment.
    - Kernel density estimation on the latitude, longitude, and time dimensions
    - Line segment intersection based approach to rank the possible PoIs
    - Requires a database containing the polygon information for each PoI
  - Distance based assignment approaches
    - Assignment of the stop to the closest PoI

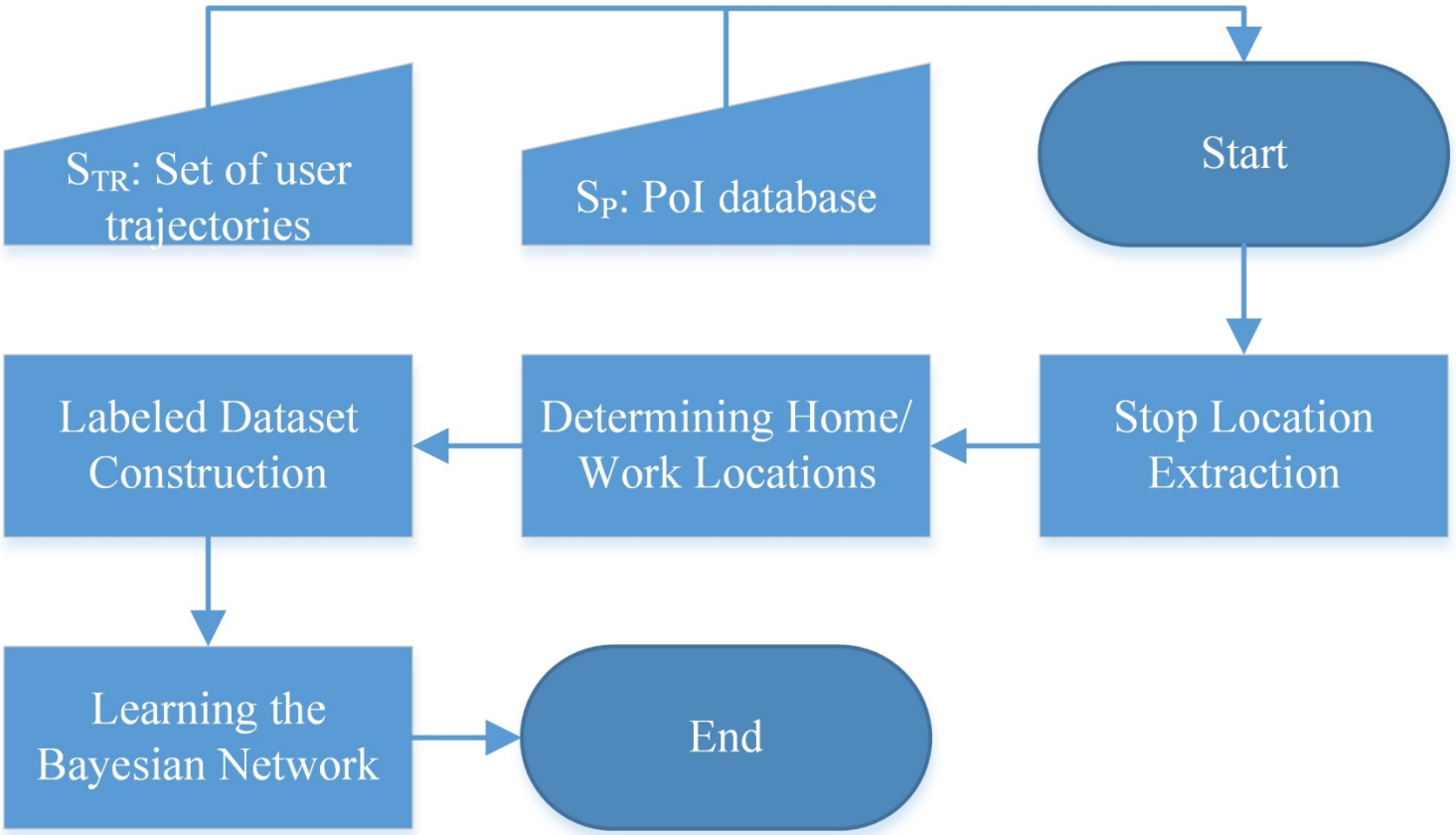
Visited PoI Extraction (VPE) Method



- Bayesian network with distance based filtering to determine the category of the visited PoI

- To learn the network, VPE includes a method to construct labeled assignment data on a subset of stops.
- In the assignment phase, the set of possible categories is the categories of PoIs within a threshold distance from the input stop.
  - The joint probability of a category and an input stop is computed, and the category with the maximum probability is the output.
  - If there is only one PoI of this category, the stop is assigned to the PoI.

VPE / Building Bayesian Network



VPE / Building Bayesian Network

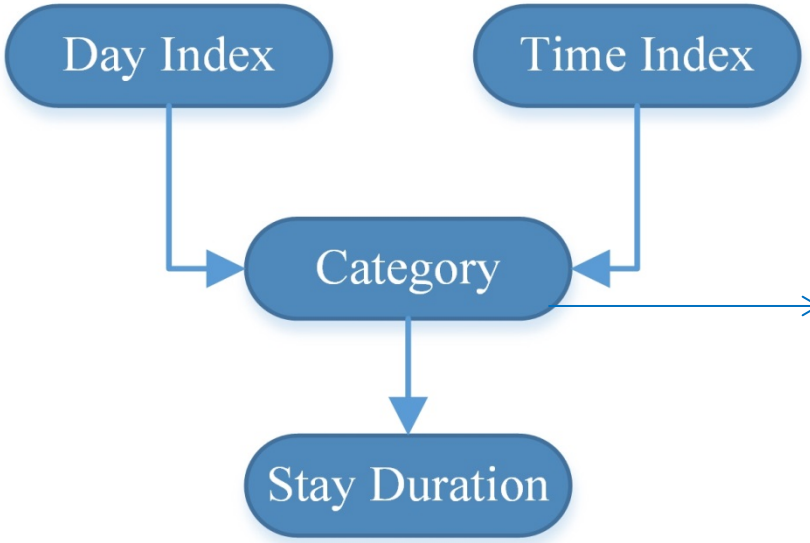
- Stop location extraction
  - Ignition mode information given by the GPS device is used.
  - If a user stops longer than a duration parameter, it is considered as a stop.
  - A distance threshold parameter is also introduced to make sure that GPS readings are correct.
- Determining the home/work locations of users
  - Density based clustering approach
    - Clustering user's stops with DBSCAN
    - If the average stay duration exceeds the input threshold (Δ_{hw}), mark all stops in the cluster as home/work stops.
  - Required in order to eliminate the visits to home and work locations

VPE / Building Bayesian Network

- Labeled dataset construction
  - A labeled dataset is needed to learn the Bayesian network.
  - This is generally not available for vehicle trajectories.
  - Distance based assignment (DBA) is used to generate labeled stops.
    - Takes a stop location and a distance threshold (ad_{th})
    - Assigns the stop location to the closest PoI if there is only one PoI inside the circular region centered at the stop location and with radius ad_{th}

VPE / Building Bayesian Network

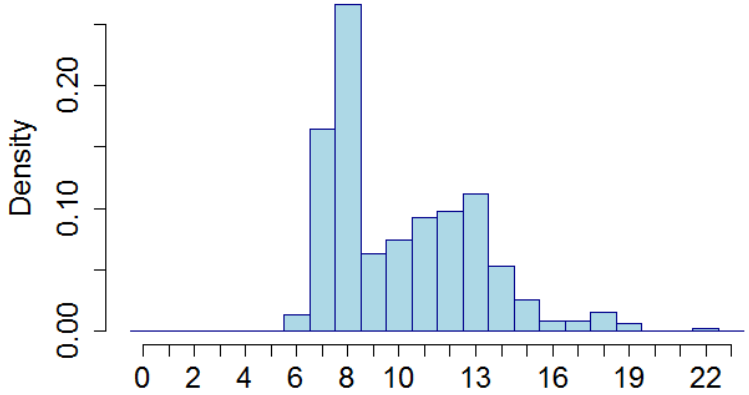
- Learning the Bayesian Network
  - Four nodes: time index, day index, stay duration, and PoI category
  - The structure of the Bayesian Network is determined according to an initial analysis on the labeled dataset

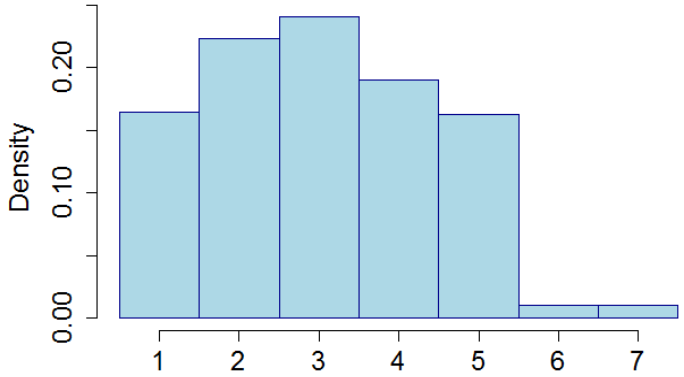


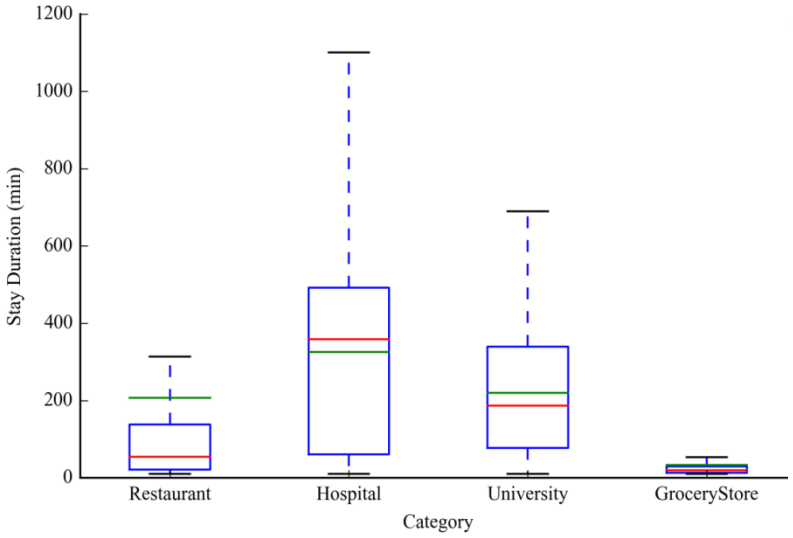
Time Index	Day Index	Category	Probability
8	1	University	0.12
8	2	University	0.23
…	…	…	…

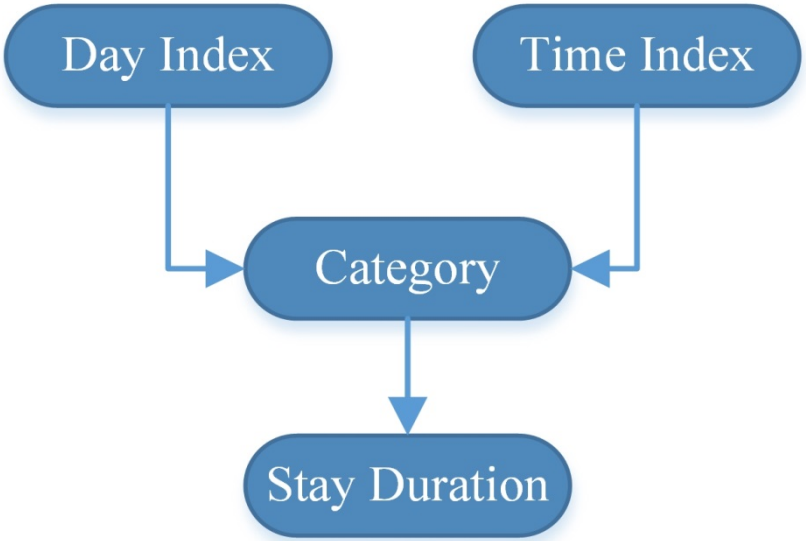
  - This step forms conditional probability tables for each node with respect to the labeled dataset.

VPE / Building Bayesian Network



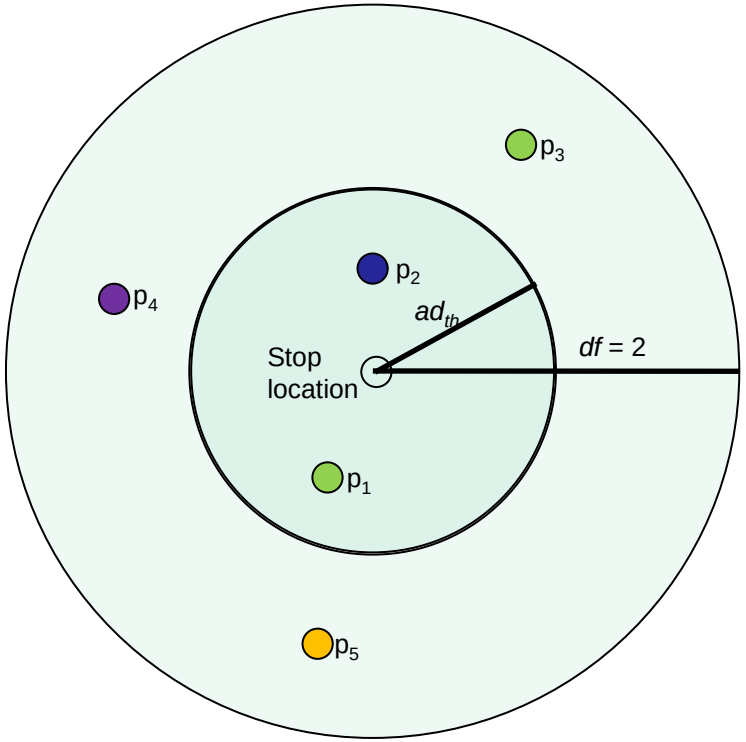






VPE / Assignment

- Distance based filtering
  - The set of possible categories is determined according to the distance factor (*df*) and the ad_{th} parameter.



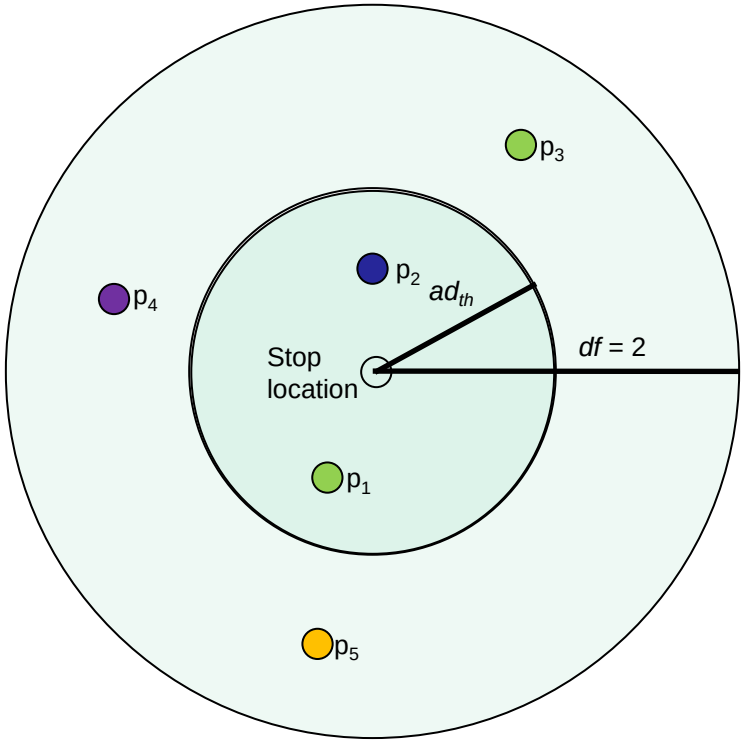
- The category of p_1 and p_3 is restaurant.
- The category of p_2 is school.
- The category of p_4 is supermarket.
- The category of p_5 is shoe store.
- So the set of possible categories is {restaurant, school, supermarket, shoe store}.

- The joint probability of the category and the stop location is computed using the Bayesian network

$$P(cat, ti, di, sd) = P(di) \cdot P(ti) \cdot P(cat \mid di, ti) \cdot P(sd \mid cat)$$

VPE / Assignment

- The category with the maximum probability is determined.
- If there is only one PoI of this category in the set of possible PoIs, the stop is assigned to this PoI.



- Assume that the category with maximum probability is supermarket.
- Then the stop location is assigned to p_4.
- If the category was restaurant, it wouldn't be possible to assign the stop since there are two nearby restaurants (p_1 and p_3).

Experimental Evaluation / Setup

- We used default values for stop location extraction and home/work stop location inference parameters.
  - Our work focuses on the *assignment* of stop locations.
- GPS data
  - 354 cars during the period 01/03/2014 – 31/12/2014
  - Contains around 0.4 billion records
  - The majority of the records are located in or around Aalborg, Denmark.
  - With the default parameters, we obtain around 350,000 stops, out of which around 130,000 are home/work stops.
- PoI dataset
  - Contains around 10,000 PoIs of 88 categories
  - Collected from Google Places
  - All of the PoIs are located in or around Aalborg, Denmark.

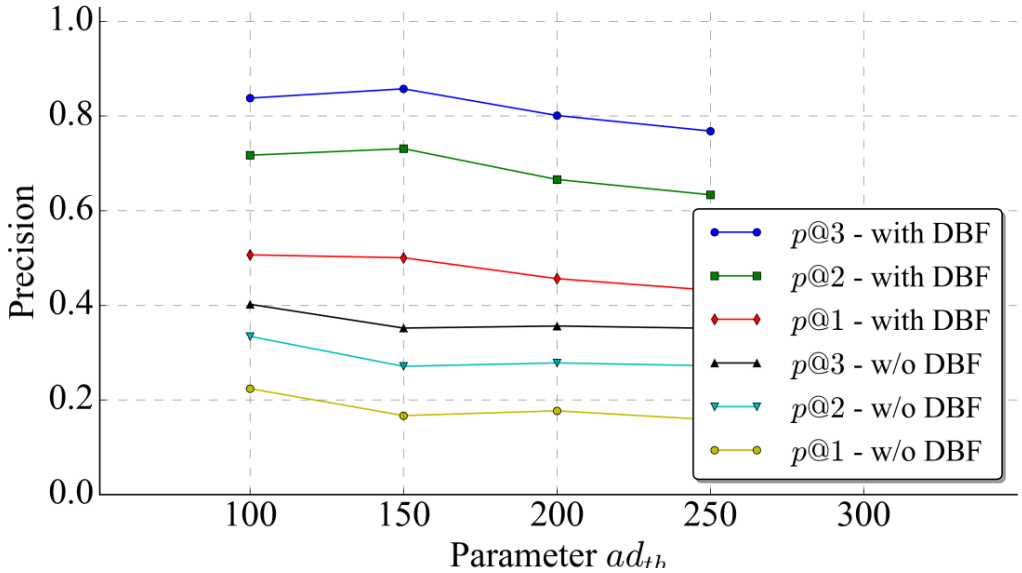
Experimental Evaluation / Setup

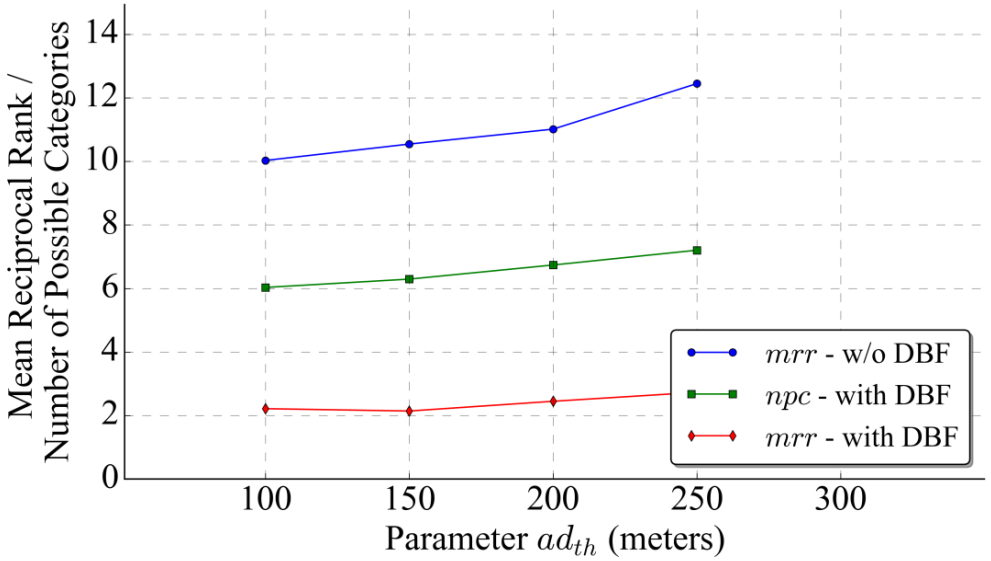
- Ground truth dataset construction for the evaluation of our assignment method
  - Labeled dataset construction as explained before with ad_{th} = 100 meters
  - Around 37,000 assignments
  - Top-5 categories are supermarket, store, school, restaurant and lodging.
- 10-fold cross validation with the ground truth dataset
- We have more than one possible PoI for each stop location in our test set.
  - We extend the surrounding area with distance factor (*df*) parameter.
  - If there is more than one PoI in this region, we add the stop location to our test set.
  - Otherwise, the stop location is not included.

Experimental Evaluation / Setup

- We modify the algorithm to return a ranked list of categories as output to evaluate the performance.
- We report the following metrics.
  - Precision at n (p@n)
    - Percentage of stops whose correct category is included in the top-n of the output
  - Mean reciprocal rank (mrr)
    - The average position of the correct category in the output
  - Number of possible categories (npc)

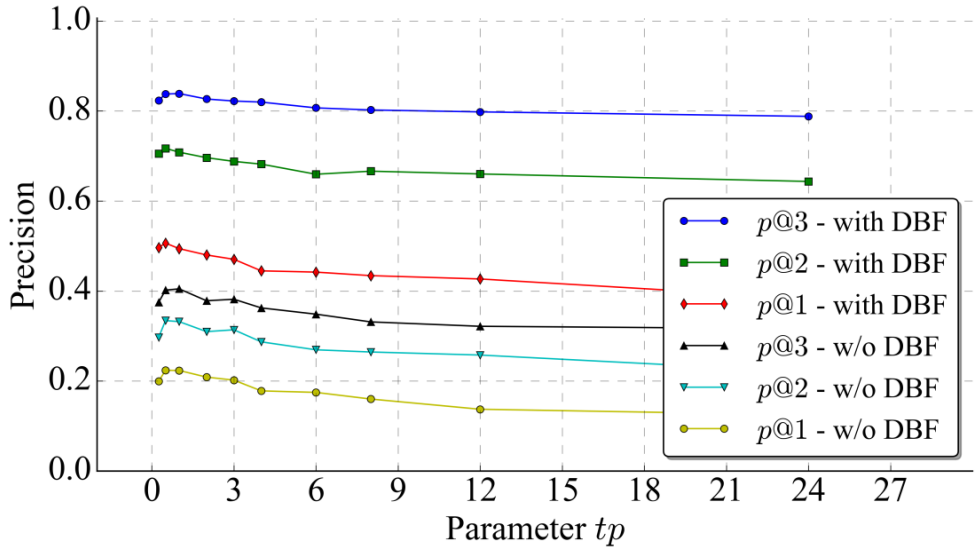
Experimental Evaluation / Results

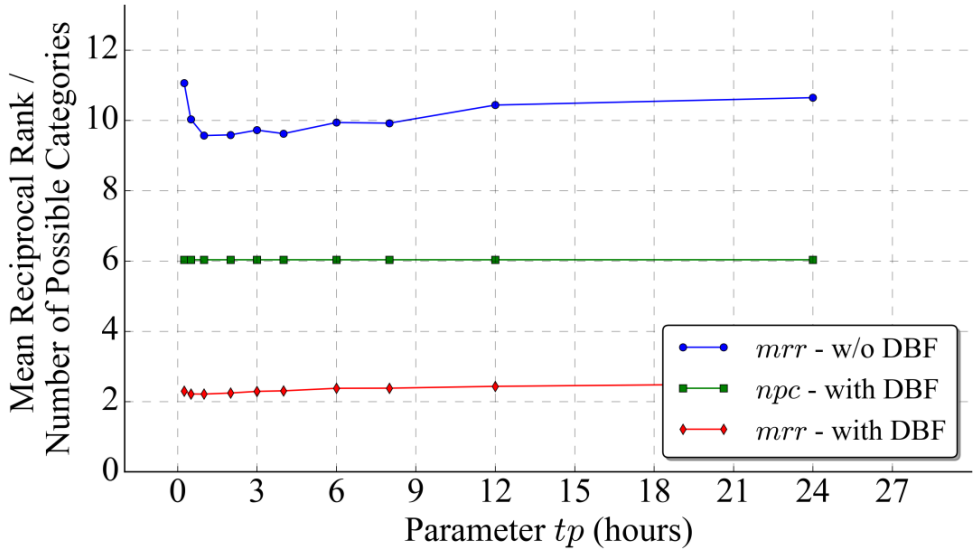




- Precision decreases when ad_{th} increases
  - The number of possible categories increases
- VPE achieves a p@3 value around 0.8 and a mean reciprocal value of 2.
- DBF has a positive effect on the precision.

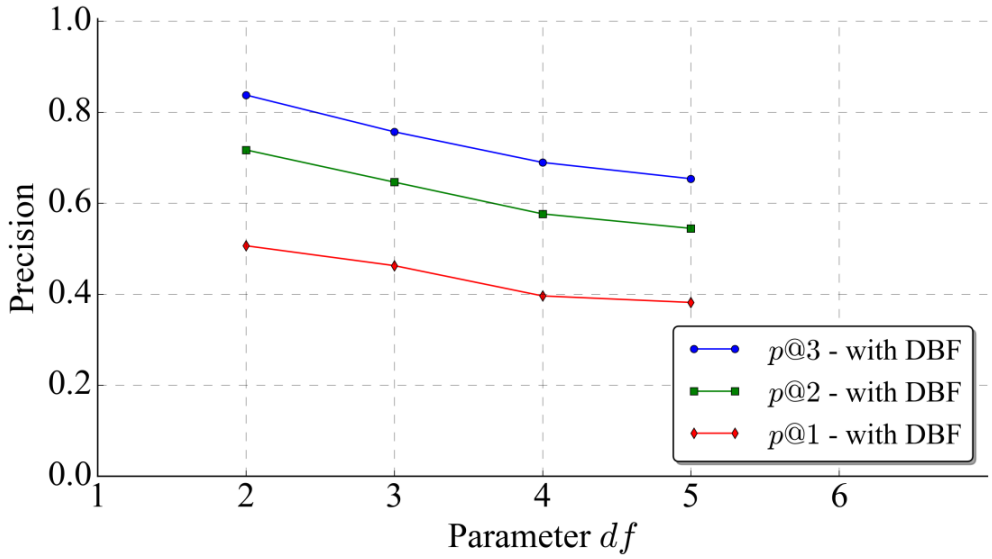
Experimental Evaluation / Results

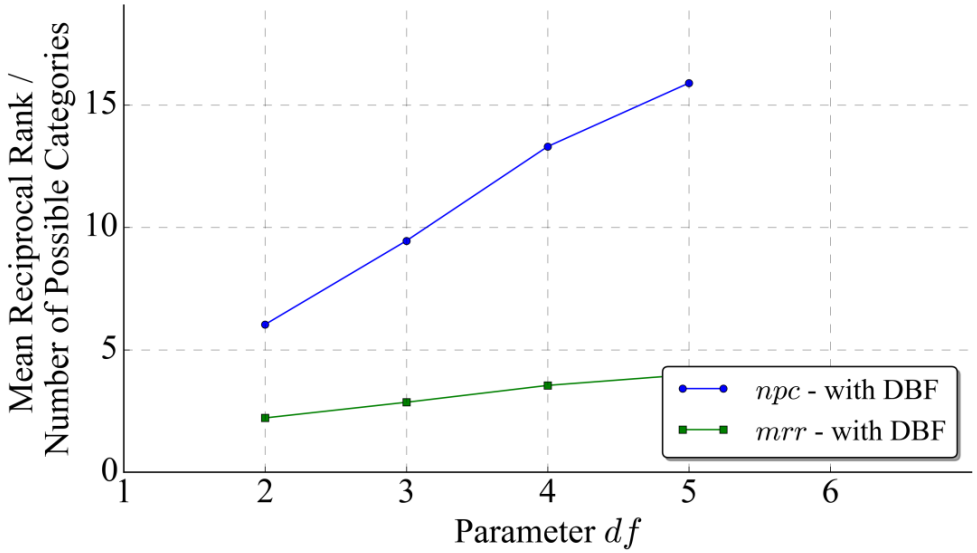




- The time period of a time slot affects the model's performance.
- The best performance is achieved when the time slot is 30 minutes or 1 hour.
- Increasing the time period decreases the model's ability to distinguish PoI categories.

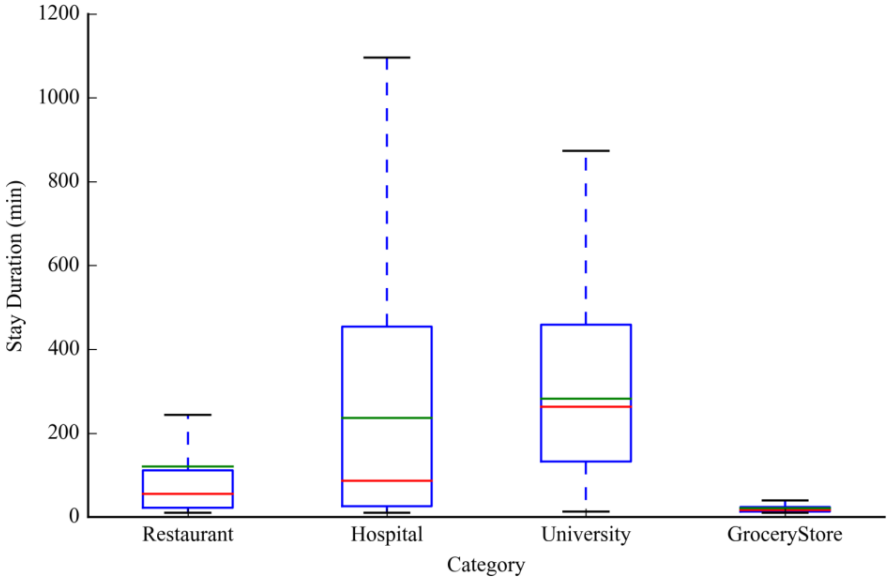
Experimental Evaluation / Results

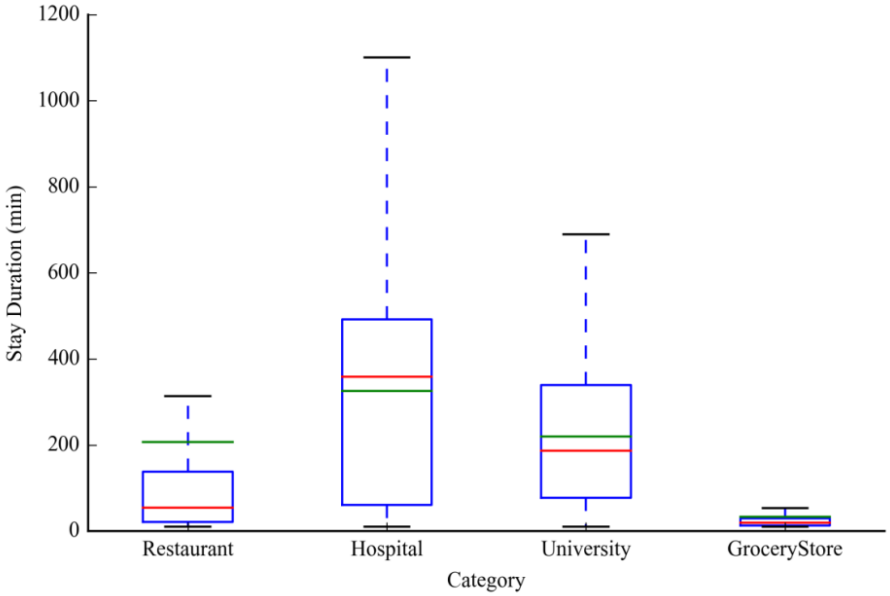




- The p@n decreases when the distance factor increases.
  - The number of possible categories increases steeply.
- The increase in the number of possible categories is steeper than the increase in the mean reciprocal rank.

Experimental Evaluation / Results





- The stay duration distribution obtained from the assignments (top) is quite similar to the one obtained from the labeled assignment construction (bottom).
- The Bayesian network is able to model the relationship between the categories and the stay duration values.

Conclusion

- We propose a Visited PoI Extraction method.
  - Employs a Bayesian network to represent the relationship between the temporal attributes of a stop and the category of the visited PoI
  - Includes a method to build a labeled dataset
- The proposed method is capable of detecting the category of the visited PoI, and it achieves a p@3 of 0.8.
- Future work
  - Combine different data sources like check-ins with GPS data
  - Use of assignment methods for evaluating ranking functions in spatial keyword queries