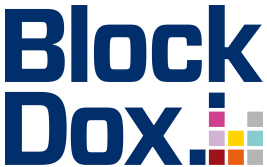
Multi-Building WiFi Fingerprinting using Bayesian and Hierarchical Supervised Machine Learning assisted by GPS

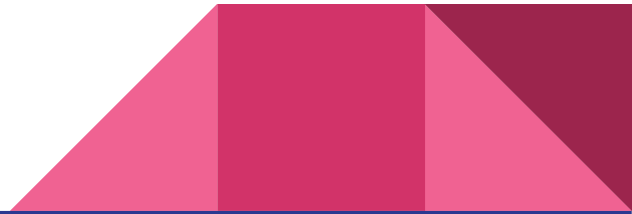
IPIN 2016, Track 3

Author: Yair Beer, Blockdox

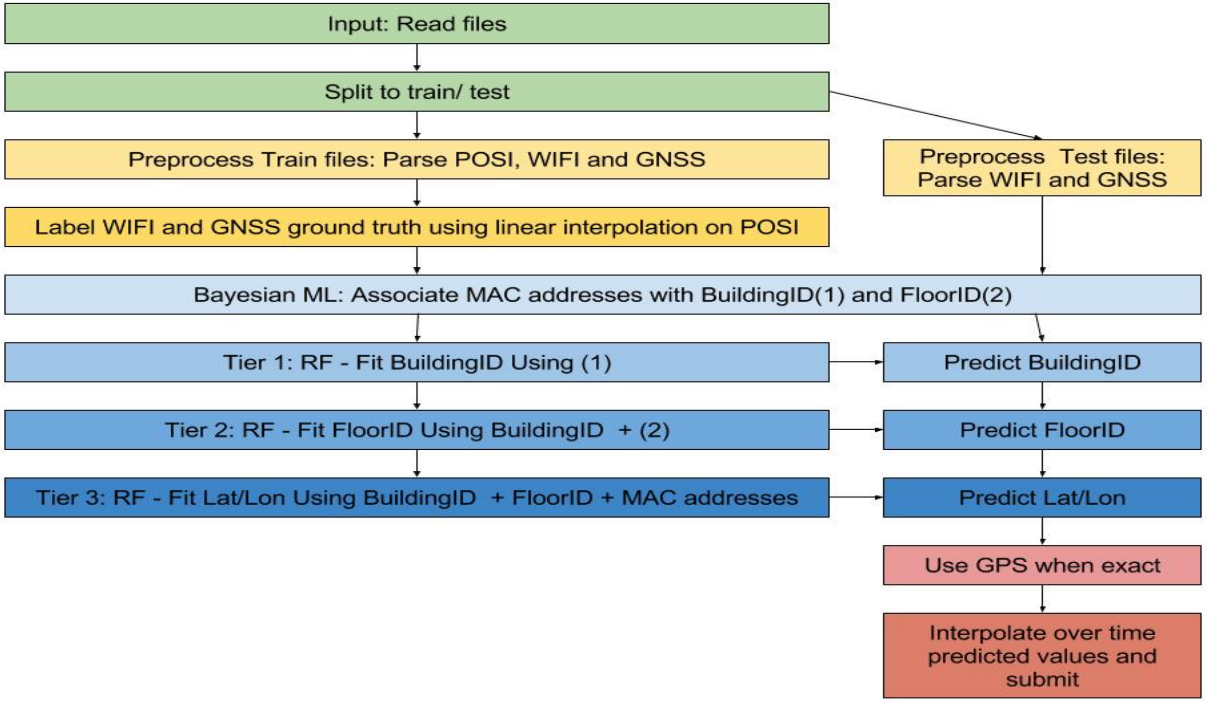


Outline

- Overview
- Bayesian Mac Address Machine Learning
- Hierarchical Machine Learning
- Cross Validation
- Time series Smoothing
- GPS Aid
- Evaluation final results
- Conclusions



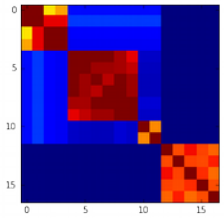
Overview


Input: Read files
Split to train/ test
Preprocess Train files: Parse POSI, WIFI and GNSS
Preprocess Test files: Parse WIFI and GNSS
Label WIFI and GNSS ground truth using linear interpolation on POSI
Bayesian ML: Associate MAC addresses with BuildingID(1) and FloorID(2)
Tier 1: RF - Fit BuildingID Using (1)
Predict BuildingID
Tier 2: RF - Fit FloorID Using BuildingID + (2)
Predict FloorID
Tier 3: RF - Fit Lat/Lon Using BuildingID + FloorID + MAC addresses
Predict Lat/Lon
Use GPS when exact
Interpolate over time predicted values and submit

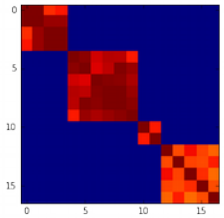
Bayesian Mac Address Machine Learning

BuildingID:

$$P(BuildingID_i) \propto P(RSSI \mid BuildingID)$$



$$P(BuildingID \mid MAC_i) \propto P(MAC_i \mid BuildingID)$$



FloorID:

- Associate MAC address with the FloorID it was measured with the highest power.
- MAC addresses measured with a power below a threshold are removed.

$$P(MAC_{floor} \mid MAC_{RSSI}) =$$

$$= \frac{P(MAC_{RSSI} \mid MAC_{floor}) P(MAC_{floor})}{P(MAC_{RSSI})} \propto$$

$$\propto P(MAC_{max\,RSSI} \mid MAC_{floor})$$

Hierarchical Machine Learning

- 3 tiers Random Forest classifier machine learning algorithm.
  - 1st Tier: BuildingID
  - 2nd Tier: FloorID
  - 3rd Tier: Latitude / Longitude
- Each tier uses the predicted result from lower tiers as features.

3rd tier - Grid Search algorithm

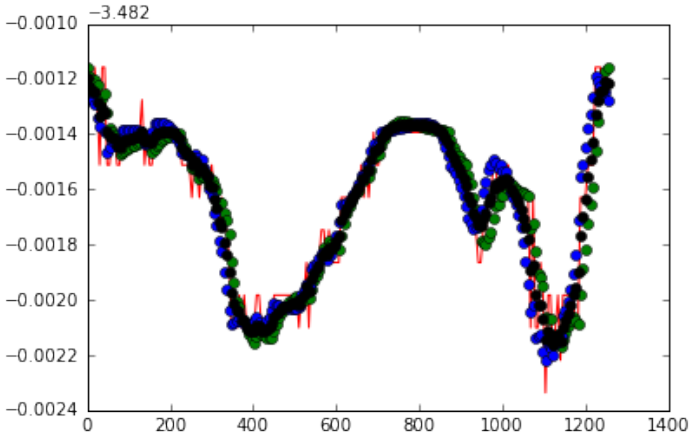
- Divide each floor and building into a grid
- Label samples into cells with corresponding Lat/Lon
- Remove empty cells
- Repeat for different grid Lat/Lon offsets

Cross validation

- Use all the routes as train data except one route.
    - If there are several runs on the same route, all of them would be used as evaluation.
- Routes:
    - 10: [0, 0, 1, 1], 20: [2, 2, 3, 3, 4, 4], 30: [5, 6], 40: [7, 7, 8, 8, 9]]
- Route 3 wasn’t use for evaluation because routes 2, 4 lacked relevant FloorID training data.
- This CV used for parameter optimization.

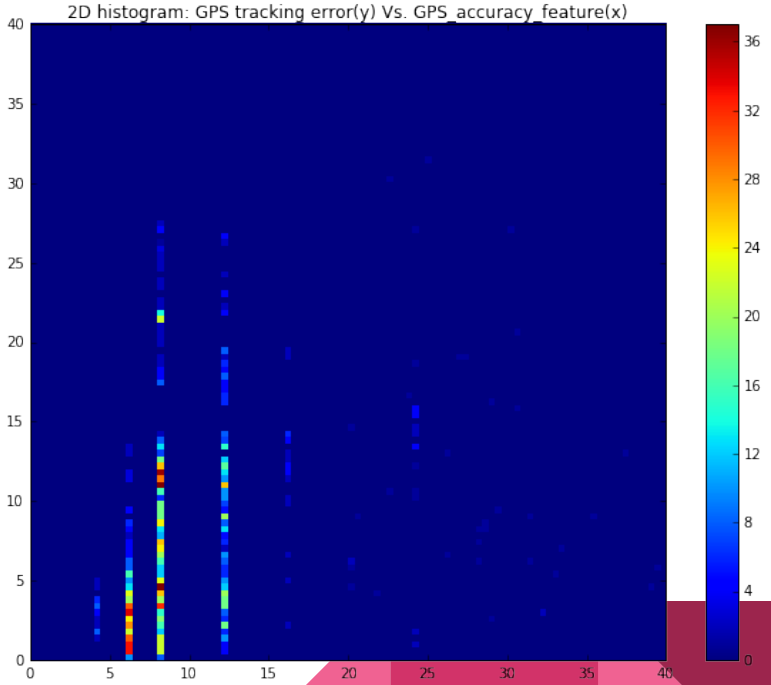
Time Series Smoothing

- Holt-Winters 2nd order exponential smoothing was used.
- When smoothing a prediction from classification:
  - The smoothing was used on the probability of prediction of each label.
- No causality restriction.
  - Averaged filtered signals from start and from the end



GPS Aid

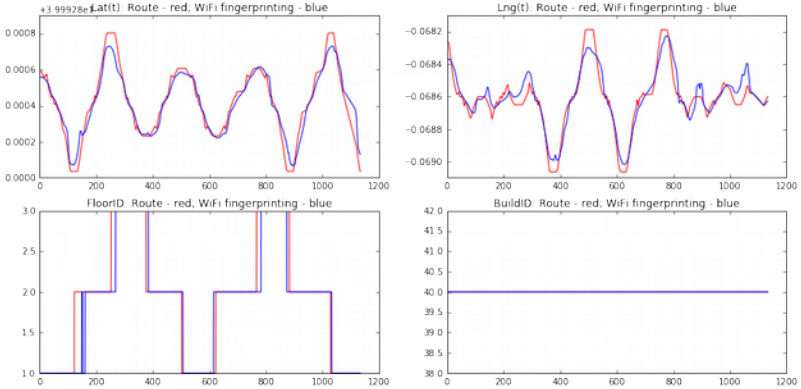
- When possible position was aided by GPS.
- The Criterion used is GPS accuracy.

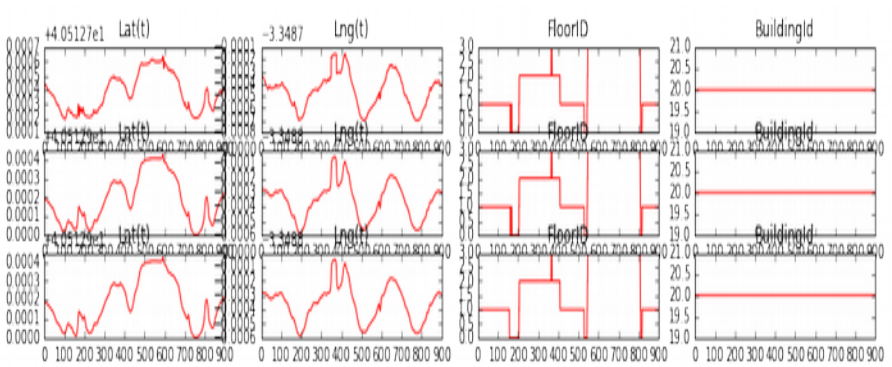


Evaluation results - MAC addresses

- Total MAC Addresses
  - 742
- MAC Addresses per BuildingID
  - 10 - 51; 20 - 353; 30 - 180; 40 - 158
- MAC Addresses per FloorID
  - 10: 0 - 39
  - 20: 0 - 190; 1 - 42; 2 - 43; 3 - 38
  - 30: 0 - 15; 1 - 16; 2 - 27; 3 - 5; 4 - 9; 5 - 57;
  - 40 1 - 48; 2 - 29; 3 - 25;

Evaluation final results - path visualisation





Conclusions

- A Robust 3 levels machine learning algorithm was introduced.
- MAC addresses association was more consistent than SSID association.
- BuildingID and FloorID associated MAC addresses reduces dimensionality and improved classification results without a priori knowledge.
- Dividing the data to routes allowed hyper parameter optimization using cross validation.
- Using GPS when reliable improved the accuracy of the measurement.

Thanks for listening.