Semi-Supervised Learning in Inferring Mobile Device Locations

Rong Duan , Olivia Hong, Guangqin Ma
rduan,ohong,gma @research.att.com
AT&T Labs
200 S Laurel Ave
Middletown, NJ 07748

Abstract

With the development of mobility technology, location information has become collectible by various positioning technologies. Different positioning technologies have their advantages and limitations. In this paper, we propose semi-supervised learning in inferring low-accuracy location data density from high-accuracy location data density. We focus on the enormous amount of low-accuracy Cell Tower Triangulation (CTT) calculated mobile device location data, and the small amount of high-accuracy Assisted Global Positioning System (AGPS) pinpointed location data. The CTT and AGPS mobile device location data is collected for each cell tower that serves the devices, then the actual distribution is learned from both CTT and AGPS data by semi-supervised learning and the likelihood for low-accuracy CTT location can be used as an accuracy indicator. The proposed method takes the advantage of the existing extensively collected location data, and augments it by a machine learning algorithm, which complements the downside of one technology with the other technology. This big data approach improves the location accuracy statistically without the added complexity and cost of upgrading or replacing mobile networks or devices. And also the proposed method focuses on the location density alignment, which avoids tracking individual user devices and preserves user privacy.
.

Keywords

Mobile Device Location, Semi-Supervised Learning, AGPS, Cell Tower Triangulation

1. Introduction

Geolocation data is essential in many fields and across many different disciplines, such as location based services, healthcare surveillance, urban planning, telecommunication capacity planning, and ecology, etc. With the development of mobile technology, the capability of locating mobile devices is achieved by different positioning technologies. Zeimpekis et al. [1] provides a good review of different technologies. In general, there are two types of positioning technologies: one is satellite position technology and the other is mobile network position technology. Satellite position technology, commonly known as Global Positioning System (GPS), is a Global Navigation Satellite System for determining positions of receivers using signals broadcasted by satellites. Currently, twenty seven satellites cover the earth with three as backup. Extended from GPS, Assisted GPS (AGPS) has been developed by Djuknic et al. [2] with an assistance server to enhance the positions performance of GPS receivers and satisfy FCC's E911 mandate requirement. Mobile network position technologies use the receiver's radio signals among multiple cell towers. Both technologies employ multilateration algorithm to determine the position of an object. Basically, the algorithm estimates locations from the intersection of multiple hyperbolic curves produced by the distance between two transmitters (satellites or Cell Towers), and the time difference it takes signals travel from the transmitters to the receivers (mobile device). This time difference is called Differential Time of Arrival (TDOA). The accuracy of the multilateration algorithm is impacted by the number of transmitters involved. More transmitters means more hyperbolas and a more precise location. Besides number of transmitters, signal strength is another important factor in position accuracy.

In general, the AGPS technology achieves very high accuracy, ranging from 1m to10m[1] since at least four satellites will be used and the signal from the satellites will travel in a more unobstructed, open space, compared to the signals traveling among cell towers. Because of the limitation of the height of cell towers, the positioning accuracy of mobile network position technologies is significantly reduced by the low Signal-Noise-Ratio. Most mobile devices will only touch at most three cell towers, Cell Tower Triangulation (CTT) is the most popular method. The accuracy of network position technologies is heavily driven by the local geographic situation, since the radio signal is very easily blocked by any vertical object. The advantage of network position technologies is that they can be applied to all mobile devices that seek mobility service, which covers the whole spectrum of mobility locations. The AGPS is relatively expensive to collect and is only available on the devices that are equipped with a GPS capability. Since not all GPS capable devices turn on the GPS function at all times, the GPS data that collected covers only a fraction of mobile devices. Overall, the AGPS data has high accuracy, but partial device coverage, while CTT data has low-accuracy, but covers all different types of devices. Fig (1) shows the map of user locations in the Florida Keys area. The red dots represent mobile user locations collected by AGPS technology and the blue dots are collected by CTT. Corresponding to our intuition, the majority of AGPS data distributes along the highway, but a fair amount of CTT data is in the water. AGPS data has a higher accuracy than CTT data. On the other hand, the amount of CTT data is more than AGPS data.

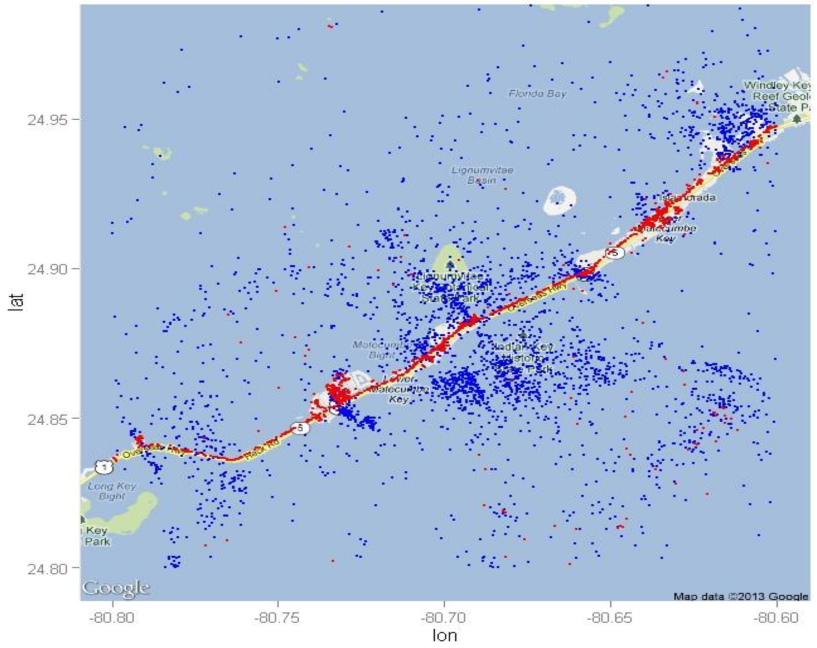


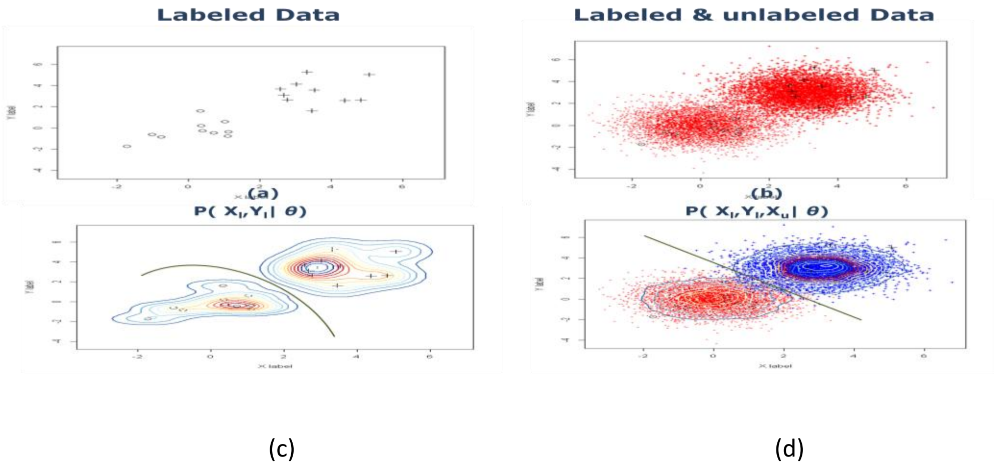
Figure (1) Location data collected by AGPS and Cell Tower Triangulation method represented by red and blue dots respectively.

Researchers have proposed different approaches to improve the accuracy of positioning mobile users. All these efforts focus on mobility techniques, signal measurements, or algorithms in improving the nonlinear math solutions. In this paper, we propose a big data approach to complement the limitations of the large amount of low-accuracy location data and the small amount of high-accuracy location data. Features that represent locations are extracted, and semi-supervised learning is adopted to gradually reveal the true location distribution base of the features. To the authors' knowledge, there has not been a study in this big data direction for improving location accuracy thus far. The advantage of the proposed method is that it improves the location accuracy statistically without the added complexity and cost of upgrading or replacing a mobile network or device. It also reveals location density information which can be used is various location related study while maintain user privacy.

This paper is organized as follows: Section 2 introduces the system and the proposed method, Section 3 evaluates the performance of the proposed method based on real world application data. Section 4 discusses the privacy issues for location based studies that are related to this paper. A conclusion and the future work are given in Section 5.

2. Approach

Semi-supervised learning has been well developed in the past few years. The general concept of semi-supervised learning is to use both labeled and unlabeled data as training data to improve the classification when the training data is limited. Zhu[3] provides a thorough review for semi-supervised learning. The generative model approach for two classes situation is as illustrated in Fig(2), (a) is labeled data with class "+" and class "o", and the whole dataset (labeled and unlabeled) shown in (b). A class boundary needed to be drawn between these two classes. If only the labeled data are used in estimating density, the decision boundary of the two classes is the curved green line in (c). The actual boundary for the whole dataset should be the line between class "+" and "o" shown in (d). The reason for the two different decision boundary is that the limited amount of labeled data can't represent the whole distribution. The generative semi-supervised learning model is to estimate the parameters base on both labeled and unlabeled data, which will get the parameters as close as possible to the true distribution.



Figure(2) Concept of the Generative Semi-Supervised Learning Model

For Gaussian Mixture Model(GMM), the joint probability for class label and input data is

$p(x,y|\ \theta) = p(y|\theta)p(x|y,\theta) = w_y N(x;\ \mu_y,\ \Sigma_y)$, where , x is input data and y is class label, w_y is mixture weights; μ_y, Σ_y are average locations and variations

The posterior probability which infers the classification label is:

$$p(y|x,\theta) = p(x,y|\theta)/\ \Sigma_{y'}\ p(x,y'|\theta)$$

Denote the labeled data as (X_l,Y_l), where X_l represents the features extracted from the original data and Y_l is the class label. X_u is unlabeled data, And assume the data follows GMM. The semi-supervised generative model is actually used to find the parameter θ that maximizes the likelihood for

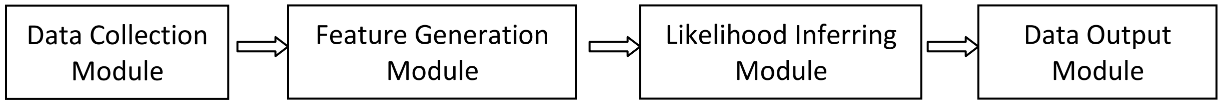
$$\text{argmax}_\theta P(\ X_l,Y_l,X_u|\ \theta) = \text{argmax}_\theta \Sigma_{Yu}\ p(X_l,Y_l,X_u\ ,Y_u|\theta)$$

The essential element of the generative GMM model in semi-supervised learning is to assign each unlabeled data to one of the mixtures and achieve the maximum likelihood for all mixtures. Each mixture is considered as one class and the unlabeled data is trained in a batch mode. Multiple researchers have observed that unlabeled data is guaranteed to improve accuracy when the model assumption is correct. But if the model assumption is not correct, unlabeled data will reduce the performance[4,5,6,7].Cozman et al.[8] provides a formal derivation regarding this.

In this study, we apply the self-training semi-supervised method with the GMM model to reveal the true distribution from the limited amount of label data and a large amount of unlabeled data. Self-training is a heuristic generative model approach for semi-supervised learning, it combines labeled data with unlabeled data sequentially and gradually update the parameters to retrieve the real distribution. The advantage of proposed method is that it can incorporate the domain knowledge in parameter control. The model parameters are initialed with the labeled data (X_l,Y_l), where X_l is the feature vector that generated from the high-accuracy AGPS location data, and Y_l is the likelihood from the GMM model. The unlabeled data X_u, which is the low-accuracy CTT data, is evaluated, and the highest confidence points are added to the training data to re-train the parameter. The procedure is repeated until the threshold is reached. The whole system and the detailed procedure are proposed in the following section.

2.1 Modules in the system

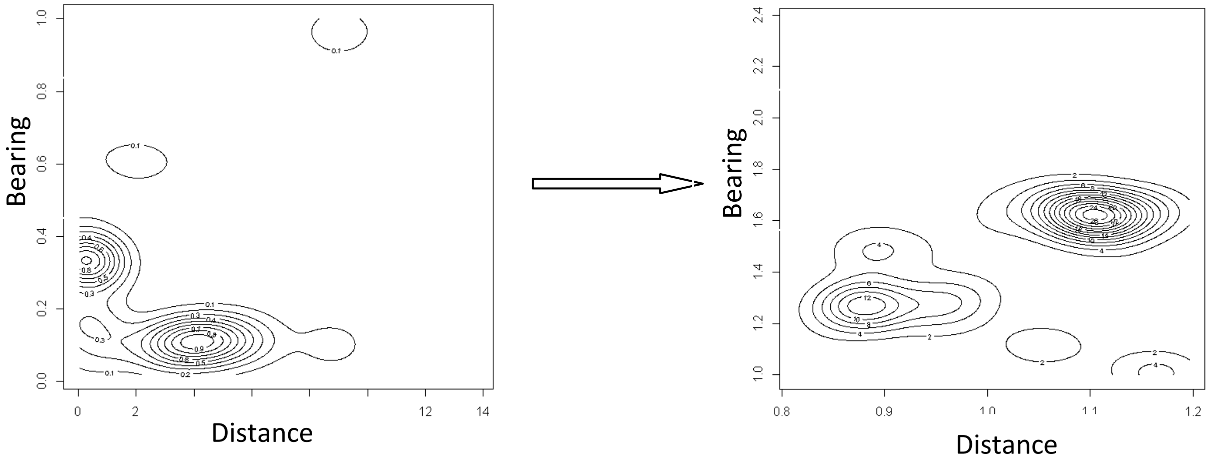
There are four modules in the system as illustrated in Fig(3): Data Collection Module, Feature Generation Module, Likelihood Inferring Module and Data Output Module.



Figure(3) System Modules

Data Collection Module: Accumulate high-accuracy AGPS and low-accuracy CTT mobile device location information, which includes the mobile device latitude, longitude and the antenna latitude, longitude that serves the device in a large scale dataset. Only the spatial location information is collected. CustomerID, timestamp, or any other information that possible to identify a customer is not collected to preserve privacy. We will address the privacy issue in detail in Section 4.

Feature Generation Module: Two features are generated to represent the device location. The distance between the antenna and the mobile device location, and the direction of the mobile device relatives to the antenna. These two features are generated based on the antenna's physical properties. One specific antenna can only cover a range of area encompassing a few directions. Since the original distance and direction might be right or left skewed, the Box-Cox power transformation is applied to adjust the input bivariate data to the bivariate normal distribution. The example is shown in Fig(4). (a) is the density estimation from the original distance and direction and (b) shows the transformed one. The original feature space has a mixture of one completed Gaussian distribution and one truncated Gaussian distribution, after the transformation, the feature space has two completed mixture Gaussian distributions.

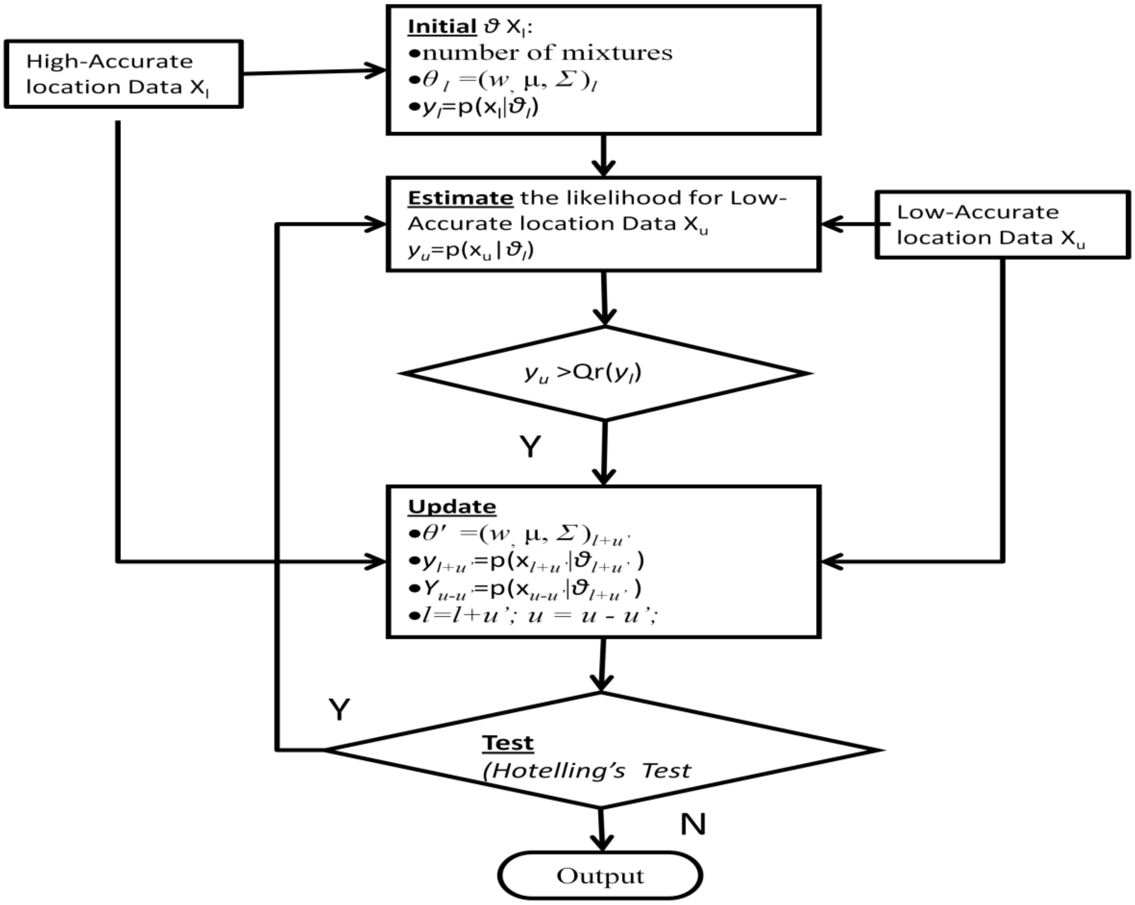


(a)Original

(b)Transformed

Figure (4) Box-Cox Power Transformed Feature Space (a) The original feature space with one completed and one truncated mixture Gaussian.(b)The transformed space with two completed mixture Gaussian.

Likelihood Inferring Module: This module is the core of the system and the flowchart is illustrated in Fig(5). There are three major steps in this module.


High-Accurate location Data X_l
Initial ϑ X_l:
•number of mixtures
•$\theta_l = (w, \mu, \Sigma)_l$
•y_l=p($x_l|\vartheta_l$)
Estimate the likelihood for Low-Accurate location Data X_u
y_u=p(x_u | ϑ_l)
Low-Accurate location Data X_u
y_u >Qr(y_l)
Y
Update
•$\theta' = (w, \mu, \Sigma)_{l+u'}$
•$y_{l+u'}$=p($x_{l+u'}|\vartheta_{l+u'}$)
•$Y_{u-u'}$=p($x_{u-u'}|\vartheta_{l+u'}$)
•$l=l+u'$; $u = u - u'$;
Y
Test
(Hotelling's Test
N
Output

Figure(5) Flow Chart for Likelihood Inferring Module

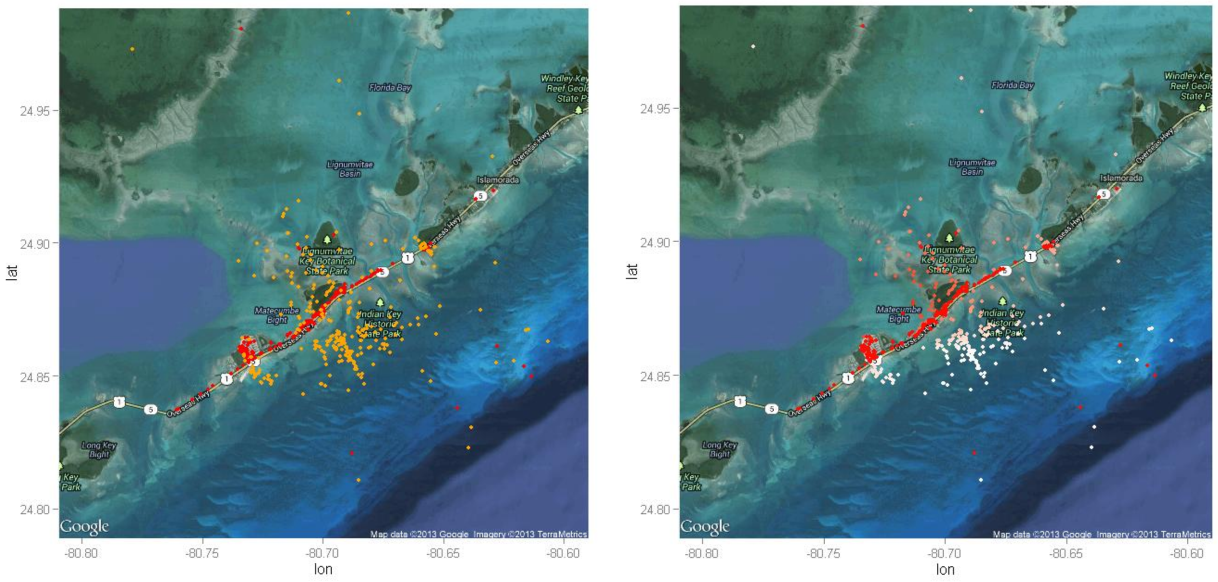
- **Initial:** The module initializes the parameters θ_l ={ w_l ,μ_l, Σ_l) with the small amount of high-accuracy location data X_l collected from the AGPS method. As indicated earlier, the critical aspect of involving unlabeled data to improve the performance is the model assumption, and any distribution can be represented as GMM if the mixture number is large enough. The important step is to estimate the optimum number of mixtures from the real data. In this approach, BIC criteria is adopted. The likelihood for X_l is y_l=p($x_l|\theta_l$)
- **Estimate:** The likelihood of the large amount of low-accuracy data X_u is estimated y_u=p($x_u|\theta_l$). Denote the location points with likelihood y_u larger than a threshold T as $X_{u'}$. The threshold T is set as the percentile of y_l, which can be adjusted by user experience.
- **Update:** Merge X_l with $X_{u'}$ as new training data $X_{l+u'}$ to update the parameters $\theta_{l+u'}$ ={ w ,μ, Σ) $_{l+u'}$ and re-evaluate the $y_{l+u'}$=p($x_{l+u'}|\theta_{l+u'}$) and $y_{u-u'}$=p($x_{u-u'}|\theta_{l+u'}$). Set the new testing data as $X_{u-u'}$
- **Test:** Compare the updated parameter $\theta_{l+u'}$ with previous parameter θ_l to control the iteration. If the difference between $\theta_{l+u'}$ and θ_l is small enough, the procedure is considered converged and the iteration is stopped. There are different tests can be adopted in this control step. Information criteria BIC/AIC is one of the options to measure the difference of the two models. We adopt multivariate *Hotelling's T Square Test*, since we believe the low-accuracy location data has larger variance comparing with the high-accuracy location data , but with the same mean.

Compared with the Generative Gaussian Mixture Model for semi-supervised learning, which optimizes parameter θ by combining X_l,Y_l,X_u and treating Y_u as missing data, this procedure uses the likelihood Y_l as a soft classification indicator and quantifies Y_u sequentially, retrains the parameters with the new set of training data and control the convergence by statistical testing, which is robust in integrating the domain knowledge.

Data Output Module: The likelihood for the original high-accuracy location data Y_l is set as the maximum of Y_u to reserve each high accuracy location point and scale to $Y \in [0,1]$. The procedure repeats for each antenna and output the likelihood for both high-accuracy and low-accuracy location data points.

3. Experiment

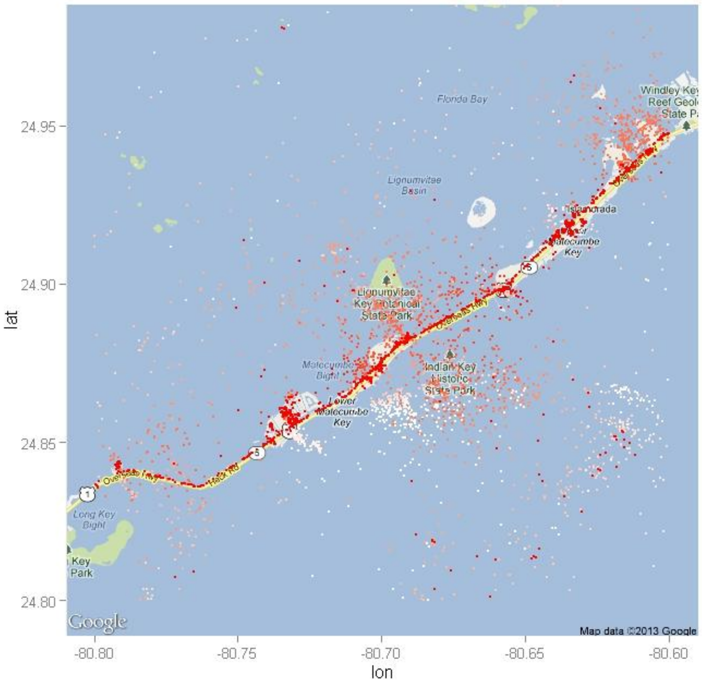
We apply the methodology on a single cell sector that is located at Islamorada, Florida Keys. The area map is shown in Fig(6) , where (a) is the original map. The red dots are AGPS data and the orange ones are CTT data. The same as Fig(1), the AGPS data is more accurate and concentrated on highway Route 1, while CTT data is spread out on the water. (b) is the likelihood map assigned by our method. The redder the color the higher the likelihood is. The CTT points on the north of the high way are assigned higher likelihood than the points on the south of the highway. We zoom in on this area and find that besides Lignumvitae Key Botanical State Park, Matecumbe Bight is also north of the highway. The north side may have more water activities than the south side.



(a) (b)

Figure(6)

Applying the method sector by sector to the area shown in Fig(1), the likelihood results are shown in Fig(7). The redder the color is the higher the likelihood.



Figure(7) The likelihood weighted location map

4. Privacy Discussion

Location data collection and usage pose growing concerns in customer privacy. The precise location tracking data can be used as biometric identifier as Jennifer Lynch and Jeff Jonas discussed at SXSW Interactive[9]. If a dataset or method could define or identify a customer, the customer privacy is at the risk to be revealed.

The semi-supervised learning method proposed in this paper only use latitude and longitude information. Any customer identification information or usage patterns are not collected. A statistical model is used to describe the spatial relations among the location data points. Even though the proposed method weighted the accuracy at the finest location point level, it still preserves customer privacy by not extracting customer location patterns or connecting the location with customer.

5. Conclusion

This paper proposes semi-supervised learning from big data to improve the location data accuracy measurement. The proposed method infers the accuracy of low-accuracy data from the existing collected low-accuracy and high accurate location data without extra network investment. The proposed method also extracts features to represent the relation between mobile user location and the served cell sector, which partitions a large scale data to a manageable cell sector level and is efficient in computing. The proposed method maintains customer privacy while retaining the detailed location information.

References

1. Zeimpekis V, Kourouthanassia PE, Giaglis GM. Mobile positioning technologies in cellular networks: an evaluation of their performance metrics. T*elecommunication systems and Technologies Vol. I Mobile and Wireless Positioning Technologies.*
2. Djuknic GM, Richton RE. Geolocation and Assisted GPS. *Communications*
3. Zhu X. Semi-supervised learning literature survey. 2006; http://www.cs.wisc.edu/~jerryzhu/pub/ssl_survey.pdf
4. Castelli V, Cover T. The exponential value of labeled samples. *Pattern Recognition Letters*, 1995; 16, 105–111.
5. Castelli V, Cover T. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory,* 1996; 42, 2101–2117.
6. Duan R, Jiang W, Man H. Semi-supervised Image Classification in Likelihood Space. *Thirteenth International Conference on Image Processing(ICIP)*, 2006
7. Ratsaby J, Venkatesh S. Learning from a mixture of labeled and unlabeled examples with parametric side information. *Proceedings of the Eighth Annual Conference on Computational Learning Theory,* 1995; 412–417.
8. Cozman F, Cohen I, Cirelo M. Semi-supervised learning of mixture models. *Twentieth International Conference on Machine Learning(ICML)*. 2003
9. Lynch J, Jonas J . I Know Where You're Going: Location as Biometric. SXSW Interactive 2013