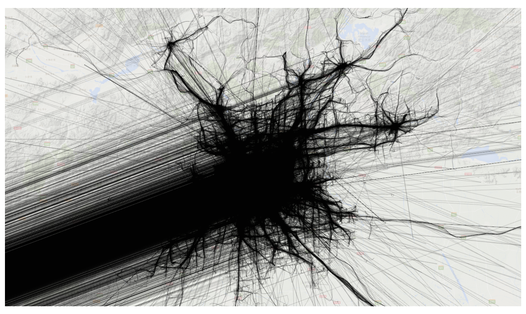
Visual Data Quality Analysis for Taxi GPS Data

Zuchao Wang[1,3] Xiaoru Yuan[1] Tangzhi Ye[1] Youfeng Hao[1] Siming Chen[1] Jie Liang[1] Qiusheng Li[2] Haiyang Wang[2] Yadong Wu[2]

1) Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University, Beijing, P.R. China
2) School of Computer Science and Technology, Southwest University of Science and Technology, Sichuang, P.R. China
3) Now at Network Security Research Lab, Qihoo 360 Co. Ltd., Beijing, P.R. China

What Quality Problems in This Data?



- Any systematic way to detect quality problems?
- Any way to detect unknown problems?
- Any way to discover strange shapes?

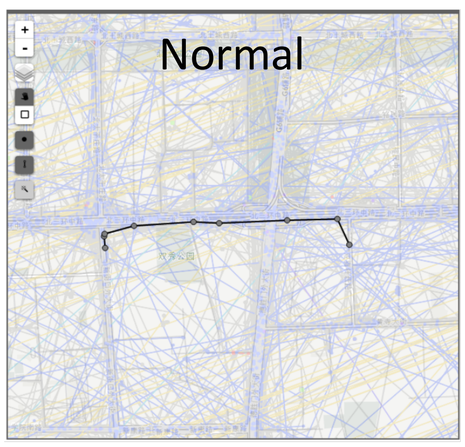
See What We Discovered in a Sample!

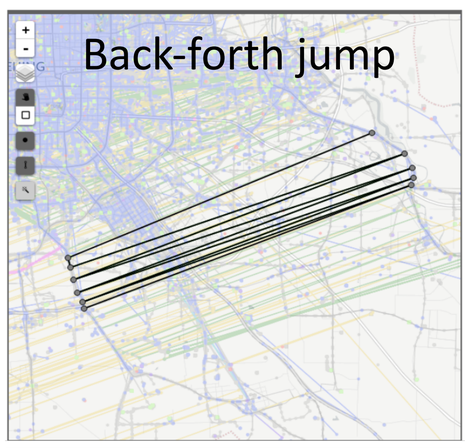
Beijing taxi GPS data

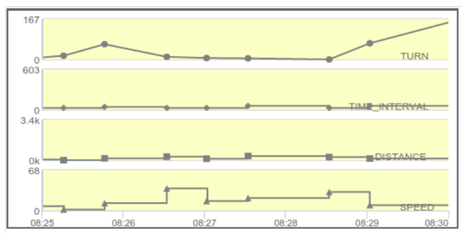
- Time: Mar 4th, 2009, 7-9 am (2 hours)
- Taxis: 13,080 (50% random sample)
- Points: 747,431
- Data size: 74.2 MB
- Sampling: 30 secs/point
- **Partition: 5 min pieces**

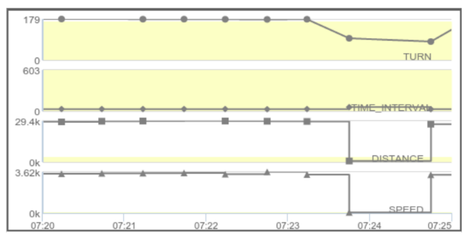
Find 8 quality problems

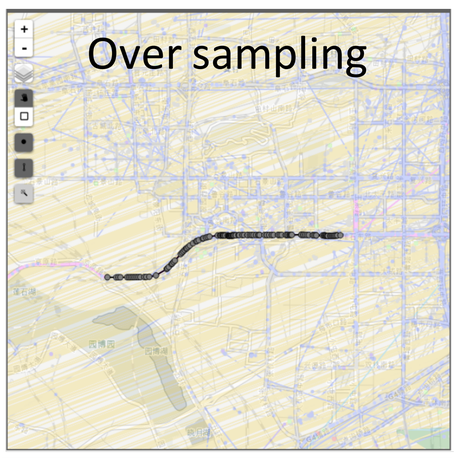
- **Over sampling**
- Data missing
- Duplication
- Stopage
- Long jump
- V-shape jump
- **Back-forth jump**
- **Random movement**

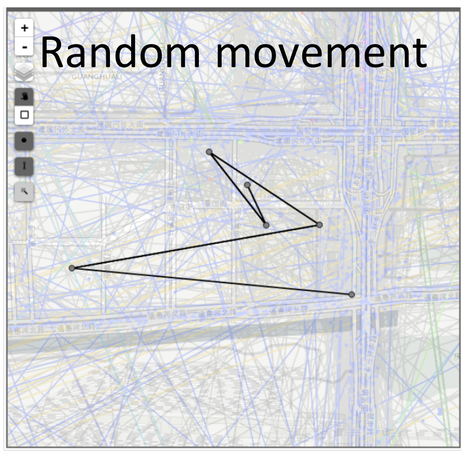


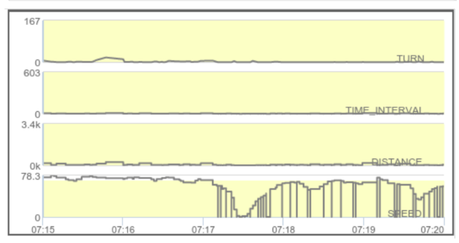


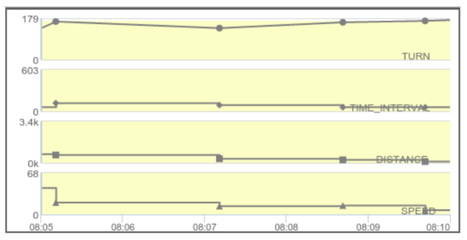












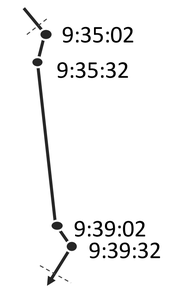
How Do We Discovered Them?

Interactive discovery of potential problems

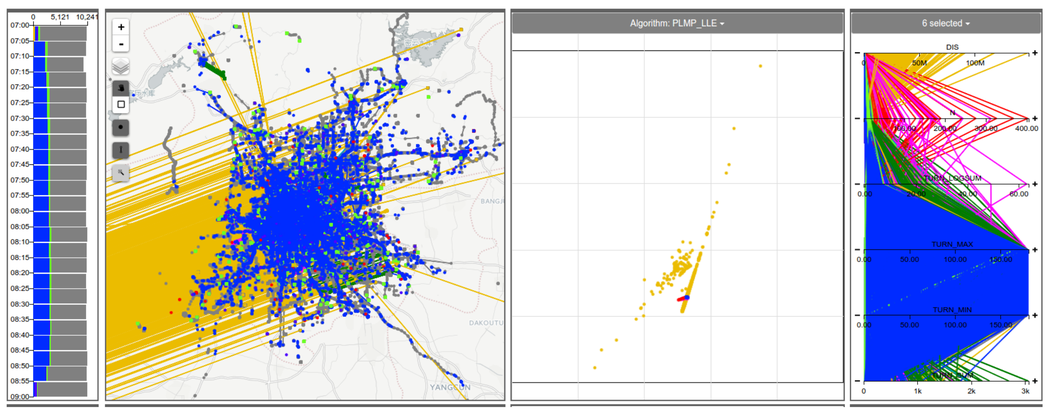
- Project trajectory pieces into spatial, temporal & **feature** spaces
- Detect data **outliers and clusters** (potentially problematic)

Feature space definition (19 features)

- #sampling points, distance, sum(turning angles)
- min/max/recsum/logsum of turning angle
- min/max/recsum/logsum of segment distance
- min/max/recsum/logsum of segment time interval
- min/max/recsum/logsum of segment average speed



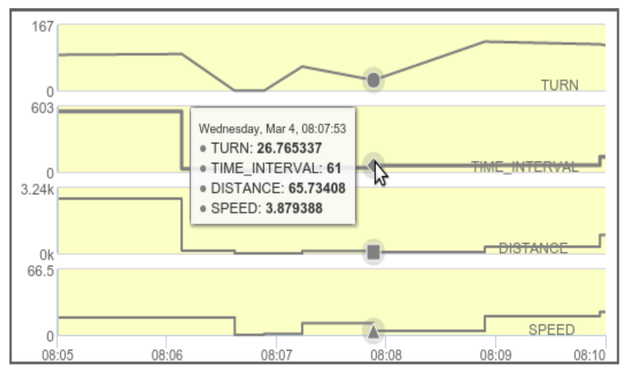
Showing data distribution from different aspects



Temporal Spatial Feature Feature

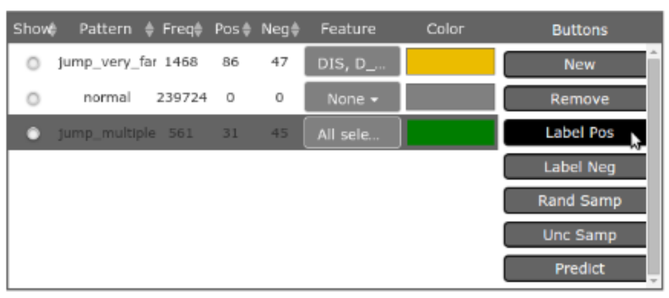
Visual confirmation of suspected data

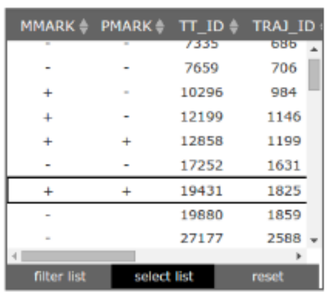
- Show the spatial shape and point/segment attribute change
- Help expert users confirm whether there's a problem



Automatic extraction of data with confirmed problems

- Binary SVM classifiers to extract problematic data
- Classifier training with interactive labeling
  - Active learning: label data that best optimize classifiers
  - Similarity search: label data similar to problematic ones





Next Step ?

- Improve scalability
- Select features systematically
- Consider interactions among different quality problems

Funding

This work is supported by NSFC No. 61170204, and partially funded by NSFC Key Project No. 61232012 and the National Program on Key Basic Research Project (973 Program) No. 2015CB-352500. This work is also supported by PKU-Qihu Joint Data Visual Analytics Research Center.

Contact

xiaoru.yuan@pku.edu.cn
http://vis.pku.edu.cn