A System for Citywide Railway Traffic Analysis Through GPS Trajectories

学籍番号　47156803
氏　　名　XIA TIANQI
指導教員　柴崎 亮介 教授

1 Introduction

Rail transportation system plays an important role in urban development as it is an energy-efficient and relatively punctual way for transporting passengers and freights. According to the Person Trip Survey conducted in 2008, almost half of the citizens in Tokyo travel by train or metro service. A serious issue in railway traffic analysis is the detection of abnormal events which result in enormous financial losses to the society. Therefore, understanding railway traffic and detecting the abnormal events is of significance to build intelligent transportation system and smart city as well as to reduce the loss from the anomalies.

With the popularization of smartphones and the development of location based services (LBS), huge amounts of GPS data generated by these LBS applications become available for traffic analysis. Over the past decades, a lot of researches have been focused on anomaly detection for road network using GPS data while few of them pay attention railway network.

Analysis of railway traffic using GPS trajectory data is challenging due to the following factors:

- Separating railway GPS trajectories with other trajectories is usually a challenging data preprocessing task.
- The topology of rail lines is more complex than roads as transferring from one line to another requires extra walking distance, waiting time and sometimes traveling cost.
- Citywide Railway traffic have spatial and temporal dependencies which are related to rail lines and timetable. For example, an accident happens in a rail line may have larger effect on passengers in the same line with a long distance than passengers in other lines closed to the accident.

Thus, in order to better extract railway traffic patterns and detect anomaly through GPS trajectories, an intelligent system is proposed in this research by using big and heterogeneous data as is shown in Figure 1. The system structure has three main modules. The data preprocessing module extracts the trajectory of railway passengers and matches them to the railway network. The pattern extraction module utilizes a kernel density estimation method to calculate the kernel density of passengers on railway network to represent the crowdedness. The anomaly detection module mainly utilizes machine learning and statistic methods to find out the segments with abnormal traffic density.

2 Related Work

2.1 Research on network kernel density estimation

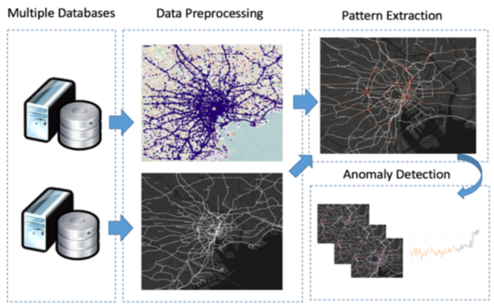

Figure 1 System Overview: the intelligent sytem utilize hetrogenous data to analyze railway traffic with three main modules.

concerning network KDE on its application, adaption and implementation. The most case that network kernel density has been applied to is the analysis traffic accidents [Xie and Yan, 2008; Okabe et al., 2009]. [Okabe et al., 2009] propose a network kernel density estimation model with two kinds of kernel functions named discontinuous and continuous function and prove that their estimation is unbiased. With the development of spatial-temporal research, KDE has been extended to spatial-temporal KDE for a more wildly use such as crime analysis [Nakaya and Yano, 2010] and network-based KDE is extended into a spatial-temporal one in [Tang et al., 2016] for trajectory analysis. Comparing to our work, this paper mainly focusses on picking up linear events instead of predicting real- time status.

2.1 Understanding human behavior with GPS trajectory data.

In recent years, several researches have focused on predicting traffic conditions using deep learning methods. [Ma et al., 2015] generalize the congestion problem as a binary classification problem and predicts traffic congestion on each road link with the energy-based deep model RNN-RBM (Restricted Boltzmann Machine). [Zhang et al., 2016; 2017] propose a grid-based method for predicting in-flow and out-flow with deep spatial-temporal residential networks. Besides traffic flow, another approach for predicting the crowdedness at a citywide level is through individual trajectory analysis [Song et al., 2016; Fan et al., 2015]. These researches deal with large volume of location history and forecast the individual movement of each passenger, thus they can be applied to a large-scale scenario such as human behavior simulation in big events or disasters.

2.2 Traffic analysis with other data sources

Besides GPS trajectory, other data sources including smart card records and route query records have already shown good performance in modeling passenger behaviors. [Ma et al., 2013; Kusakabe et al., 2010; Zhou et al., 2016] In comparison to the data source of GPS trajectory, these data sources have less noise and have more precise origin-destination information. However, they also have some disadvantages, e.g. smart card records cannot predict the appropriate rail line which a passenger chooses in a complex railway network system and route query record may have a bias to collect data from the passengers who are not familiar with some specific rail lines.

3 Data Source and Data Preprocessing

In order to forecast railway traffic through GPS data, we collect data from heterogeneous databases and process these data with different approaches. The data can be listed as follows.

3.1 Railway network

Railway network data of Tokyo 23 wards and some parts in adjacent area is collected in this research. The total length of railway in research area is about 1100 km over more than 1000 stations. Then railway network is represented by a graph *G(V, E)* where V is the set of railway stations and E is the set of railway links. For the shortest-path calculation in this network, a multi-criteria shortest-path algorithm proposed by [Disser et al., 2008] is utilized so as to take transfer and company information into consideration.

3.2 GPS trajectory data

A huge volume of raw GPS records from August 1, 2010 to July 31, 2013 is collected for our research. Railway trajectory data is extracted by the following steps: firstly, we filter the whole data by a spatial buffer and then use a Hidden Markov Model(HMM) based map matching method [Newson and Krumm, 2009] to further refine the railway trajectory data. Since GPS signal is sometimes not available for subway passengers, a five-minute time base linear interpolation through network is conducted for research use.

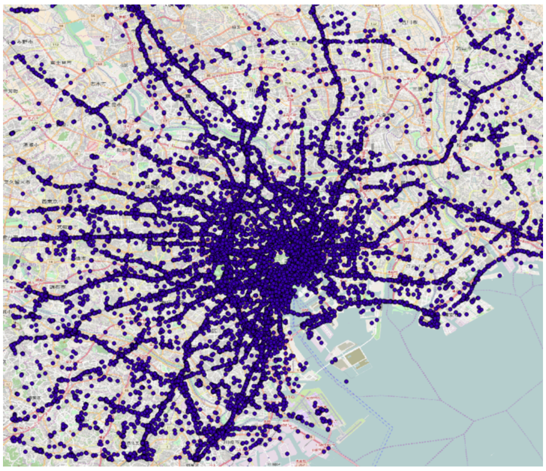

Figure 2 Visualization of Railway Trajectory Points

The calibrated GPS trajectory points are visualized in QGIS as Figure 2 shows.

3.3 Railway passenger volume data

In order to validate GPS data and map matching method. Railway Passenger Volume data is collected from National Land Numerical Information to compare with the map matching results in this research. The regression plot of passenger volume data and map matching result is shown in Figure 3.

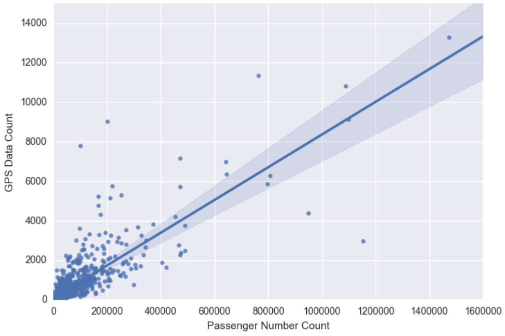

Figure 3 Regression Plot of Traffic Volume Data and Map Matching Result

4 Network Kernel Density Estimation

Network kernel density estimation(KDE) is an adaption of standard kernel density estimation which is used for calculating the density of point events over network space. [Xie and Yan, 2008]

In network KDE, the railway network is separated into segments with equal length and the density is calculated through the railway network with a kernel function which is used for measuring distance decay effect. For each kernel, only the GPS points within a bandwidth are taken into consideration.

In this research, the length of segments and bandwidth are respectively set as 300m and 1000m. Network kernel density is calculated every five minutes for the whole railway segments and the average kernel density of each thirty minutes is calculated in each day from 6:00 to 23:00. The visualization of Railway Kernel Density using QGIS is shown in Figure 4.

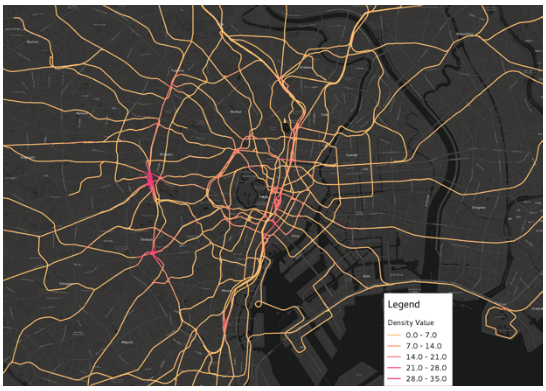

Figure 4 Railway Density in Tokyo Area

The weekly kernel density values are plotted to visualize the periodical patterns in Figure 5.

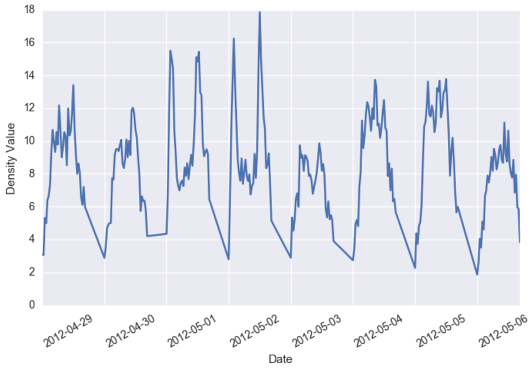

Figure 5 Density Visualization for One Segment

5. Anomaly Detection

In this research, two types of anomaly detection methods are tested in several datasets to find out the feasibility of anomaly detection with kernel density value.

5.1 PCA based anomaly detection

PCA based anomaly detection Decomposes the dataset into several principle components and then divides the axes into the components of normal traffic and abnormal traffic based on threshold. Then the principle components representing normal traffic will be inversely transformed to compare with the observed value. Figure 6 shows the anomaly detected in Comiket using PCA.

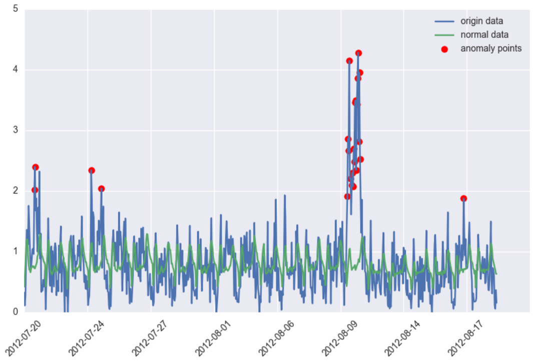

Figure 6 Anomaly Detection Example

5.2 prediction model based anomaly detection

In this research, a sequence prediction model named Seasonal ARIMA is utilized to detect the anomalies in kernel density data. At first the model is fit by training data and then the model is used to predict the sequence data. the anomaly is detected by determining whether the observed value is in the prediction interval. The data out of the prediction interval will be regarded as the anomalies. The anomaly detection results indicate that our system is feasible to detect the anomalies of big events while the performance is poor on accidents. The causes are discussed in the thesis.