Detecting Illegal Pickups of Intercity Buses from Their GPS Traces *

Ling Yin, Jinxing Hu, Lian Huang, Fan Zhang, and Peng Ren

Abstract—In some regions, especially in developing countries such as China and India, intercity bus drivers would like to pick up passengers outside the scheduled bus stations and pocket the bus fares. Such illegal pickups bring large safety risks. The alert functions offered by the current vehicle tracking systems barely can catch this illegal act, and few studies of detecting driving anomaly from GPS data focus on it either. This study hereby presents an initial effort to fill the gap. We propose an approach to automatically detecting suspicious pickup locations from intercity bus GPS traces, and implement the approach in a geographical information system. A case study demonstrates the effectiveness of the system with its high accuracy of detecting illegal pickup locations, and its functionality to help traffic police to understand illegal pickup behavior and plan a site investigation.

I. INTRODUCTION

As one of the most common methods of public transportation, intercity bus services are of prime importance in many countries and regions. An intercity bus carries passengers travelling significant distances between different cities, towns, or other populated areas. Unlike an inner-city transit bus service, which has frequent stops throughout a city or town, an intercity bus service generally has a single stop at one station in or near a city, and travels long distances without stops till next city station. However, in some countries or regions, especially in developing countries such as China and India, intercity bus drivers would like to pick up passengers at random or selected locations outside the scheduled bus stations and pocket the bus fares. Since it costs lower than regular bus fare and boarding locations could be selected, a number of passengers would choose to board outside the bus stations. It seems like a win-win situation for bus drivers and some passengers, but gradually black markets of illegal intercity bus fares are growing in size and complexity. This however is against public transport regulations and is an illegal act. Such illegal pickups act brings large safety risks. The extra passengers boarding outside bus stations would cause bus overload, and probably carry hazardous items (e.g., combustible and explosive materials) since these passengers do not undergo the safety inspections at bus stations. Overload and hazardous items are the major contributory factors that cause bus accidents. Frequently stopping largely harms passenger comfort for those who legally board at bus stations, and may lead to bus delays or augmented road congestion.

Therefore, this illegal pickup behavior has to be controlled and stopped. Police Department in many cities of China put in a number of traffic enforcement units to investigate the illegal pickup locations and try to arrest bus drivers who are conducting illegal pickups on site. An effective site investigation requires long-term experience from traffic police and takes large time and labor costs, which call for assistance from modern technology of intelligent transportation systems (ITS).

This study develops a geographical information system for traffic police (GIS-TP) to automatically detect suspicious pickup locations from intercity bus GPS traces. This type of application has not yet been found in related literature. Based on the near-repeat pattern of illegal pickups, the GIS-TP first uses a kernel density model to identify all stop clusters as a pool of suspicious pickup locations, and filters out unreasonable locations based on the cluster patterns of reasonable stops along with a series of spatial analyses. It also evaluates the suspicious level of each suspicious pickup location by analyzing its neighboring built environment and pickup behavior. The GIS-TP allows traffic police to visually examine each suspicious pickup location on map with the characteristics including suspicious level, total stops, most likely pickup hours, and highly suspicious bus licenses. Such knowledge can improve the efficiency of site investigation.

The remainder of this article consists of six parts. 1) Section 2: a review of related work on anomaly detection from GPS traces; 2) Section 3: an overview of the GIS-TP design; 3) Section 4: a proposal of the approach to detecting illegal pickup locations; 4) Section 5: a case study and a discussion of detection results; 5) Section 6: a demonstration of the web-based user interface of the GIS-TP; 6) Section 7: a summary of contributions and future work.

II. RELATED WORK

In recent years, GPS-based vehicle tracking systems (VTS) gradually become necessary equipment on intercity buses. These VTS can monitor bus speed, idle duration, route, and other observations derived from GPS traces [1-2]. The major automatic alert functions to detect driving irregularities in current VTS include speeding alert, route deviation alert, and overtime stop alert [3-4]. These alert functions are based upon simple comparisons between direct observations from GPS data and regularities such as speed limit, scheduled route, and scheduled stop time. However, intercity buses often pick up extra passengers along the scheduled route, which would not trigger the route deviation alert; extra passengers often pocket on buses within a few minutes, which would not trigger the overtime stop alert. The current VTS thus can barely help to catch the illegal pickups.

Researchers have tried to use data mining techniques to support more accurate and intelligent anomaly detections from...
GPS traces. [5] developed a conditional random field model to detect anomaly in speed, drive direction, and possible traffic jams from GPS traces. [6] and [7] respectively proposed statistical models to monitor vehicle speed violation based on GPS traces. [8] and [9] used taxi GPS traces to detect taxi driving fraud indicated by long-distance detour. [10] also used taxi drivers’ detour behavior to infer traffic jam. [11] mined the movement patterns from personal GPS data to detect disorientation of elders. These proposed anomaly detection methods neither help to detect the illegal pickups. [12] and [13] designed solutions for stay point detection from GPS traces. Their works however do not infer if the stay points are illegal pickup locations or not.

This study is an effort to fill the gap to assist in detecting illegal pickup locations of intercity buses from their GPS traces.

III. OVERVIEW OF THE GIS-TP DESIGN

GPS traces from intercity buses serve as an input to the GIS-TP. A GPS data item consists of five attributes: (ID, latitude, longitude, speed, and time). ID is the unique license number of an intercity bus. Latitude and longitude define the geographical location of the bus at the time of data collection. Speed is the velocity of the bus at the time of data collection. Time is the time stamp of the GPS data item.

![Diagram of GIS-TP Design](image_url)

Figure 1. The overview of system architecture

As Figure 1 illustrates, the web-based user interface forms the front-end of the system, which will be demonstrated in the Section 6. The back-end consists of six major components: 1) A GPS data preprocessing module: it deletes data items with null values. 2) A stop event detection module: it identifies all stop events of intercity buses. A stop event is defined by at least two consecutive data items with zero speed. The latitude and longitude of data items within a stop event are not always identical due to GPS errors. This module defines the location of a stop event by the average latitude and longitude of data items within the stop event. It also calculates the duration of a stop event. Because most illegal pickups only take a few minutes, we set a stop duration threshold \( T_{stop} \) to reduce the pool of suspicious stops. 3) An illegal pickup detection module: it infers suspicious pickup locations based on the stop events. This module will be discussed in detail in the Section 4. 4) A cloud computing platform: it is implemented by a multi-node Hadoop cluster, and supports the GPS data preprocessing module and the stop detection module. 5) A GIS engine: it offers spatial analysis functions, and supports the illegal pickup detection module. 6) A GIS sever: it provides web services for the detection results and maps.

IV. ILLEGAL PICKUP LOCATION DETECTION BASED ON GPS TRACES

We design the approach to detecting illegal pickup locations based on typical illegal pickup patterns and the investigation focus from traffic police. Many intercity buses tend to choose the same or near pickup locations to attract more passengers. Such pickup locations will become fixed and public ones in the black market. When the market grows to a certain size, it might be discovered by traffic police. Then these pickup locations will be changed. This illegal pickup pattern falls clearly the near-repeat phenomenon in criminology, which states that if a location is the target of a crime, then locations within a relatively short distance have an increased chance of being targets for a limited period of time [14-15]. Besides, traffic police focus on these pickup locations with a certain size instead of occasional and random ones. Therefore, the first step is to detect stop clusters with high density of stop events within a period of time.

**Step 1: Detecting Stop Clusters.**

To efficiently identify stop clusters and accurately extract the boundary of a spatial cluster, we use the kernel density estimation to generate a density map of stop events within a given period of time. Kernel density estimation is a statistically non-parametric way of inferring the probability density function of a population based on a finite data sample [16]. It is commonly used in spatial analysis to calculate a magnitude per unit area from point features using a kernel function to fit a smoothly tapered surface to each point. Because observed bus stop locations within a period of time can be considered a finite point data sample for calculating the probability density of all stop events, kernel density estimation is one suitable method to serve our objective. Specifically, we partition the target study area into a rectangular grid with spatial resolution \( L_{grid} \) and calculate density probability for each grid block by the kernel density estimator (Equation 1–3).

\[
\hat{f}_h(x, y) = \frac{\Sigma_{i=1}^n K(d)}{nh} 
\]

(1)

\[
K(d) = \begin{cases} 
\frac{3}{4} (1 - d^2), & d \in [0,1] \\
0, & otherwise 
\end{cases}
\]

(2)

\[
d(x,y) = \sqrt{(x-x_0)^2 + (y-y_0)^2} 
\]

(3)

where \((x,y)\) is the centroid of a grid. \(K(d)\) is a kernel function, \(d\) is the distance to a grid, \(h\) is a smoothing parameter called the bandwidth, \(n\) is the number of grids with a \(d\) less than the bandwidth of \(h\). The \(K(d)\) is a quadratic kernel function [16].
Based on the density map of stop events, a density threshold $DEN_{stop}$ is used to identify grid blocks of high density. Connected grid blocks with high density form a single stop cluster. The accurate spatial shape of a stop cluster is created by smoothing the outer boundary of these connected grid blocks (see examples in Figure 2). The density threshold is customized according to the requirement of traffic police.

**Step 2: Inferring Suspicious Pickup Locations**

Identifying suspicious pickup locations from all stop clusters $R_{stop}$ is the most challenging step of this study. When an intercity bus travels, it may stop because of various reasons, such as stopping at tolls, waiting for traffic signals, blocking by traffic jams, gas refueling, vehicle maintenance, taking a rest, and so on. After investigation on hundreds of stop clusters, we propose an elimination process to infer suspicious locations by recognizing reasonable stop patterns. The four typical patterns of reasonable stops include:

Pattern 1: The scheduled stops. Each scheduled stop has a stop region represented by a polygon $R_{station}$. The boundary of a $R_{station}$ is determined based on the real area of a bus station on a satellite image. The elimination process will exclude $R_{stop}$ spatially intersecting with $R_{station}$ (Equation 4).

$$R_{stop} \cap R_{station} = \emptyset \quad (4)$$

Pattern 2: The stop clusters waiting for traffic signals. Each traffic signal is represented by a point $P_{signal}$. The elimination process will exclude $R_{stop}$ whose minimum distance to a $P_{signal}$ is less than a distance threshold $D_{signal}$ (Equation 5).

$$\min \left( \text{Dis}(p^{j}_{\text{stop}}, P_{signal}) \right) < D_{signal} \quad (5)$$

where $p^{j}_{\text{stop}}$ is the points composing the boundary of a $R_{stop}$, $\text{Dis}()$ is the Euclidian distance.

Pattern 3: The stop clusters blocked by traffic jams. Since the traffic queue caused by congestions often has a long and narrow shape along a road, we call this shape strip in this study. The elimination process will exclude $R_{stop}$ with a strip shape and an average travel speed less than a speed threshold $V_{traffic}$ (Equation 6). The shape of a stop cluster is measured by a shape index $SI$ [15]. The minimum value of $SI$ is 1, which indicates a circle. The larger the $SI$ is, the less similarity with a circle the shape has. This study defines a strip whose $SI$ is larger than 1.2, which indicates a rectangle with a longer side twice of the shorter side.

$$SI = \frac{c}{2\sqrt{AA}} > \text{SI}_{\text{strip}} \text{ AND Avg\, speed} < V_{traffic} \quad (6)$$

where $C$ is the circumference of a shape, and $A$ is the area of the shape.

Pattern 4: The other reasonable stops. The other reasonable stops include tolls, checkpoints, auto repair shops, and auto care shops. Each reasonable stop is represented by a region $R_{other}$. The boundary of a $R_{other}$ is also determined based on the real area of a facility on a satellite image. The elimination process will exclude $R_{stop}$ spatially intersecting with $R_{other}$ (Equation 7).

$$R_{stop} \cap R_{other} = \emptyset \quad (7)$$

After the elimination process, the left stop clusters are the suspiciously illegal pickup locations.

**Step 3: Evaluating Suspicious Level**

We propose three suspicious levels for suspicious pickup locations. Intercity bus drivers tend to pick up passengers near some particular facilities including intercity bus station, transit bus station, metro station, parking lot, gas station and locations of travel agents. Suspicious pickup locations near these facilities have the highest suspicious level. Except the above mentioned locations, if a suspicious location has repeated stop events of a bus, it has a secondarily suspicious level. The rest other locations have a low suspicious level.

In order to offer site investigations more knowledge, the GIS-TP also calculates most likely pickup hours and highly suspicious bus licenses for each suspicious pickup location. Assuming stop events have a Poisson distribution over time (Equation 8), the probability of occurring at least one stop event in a time interval can be calculated to define the most likely pickup hours. The users can adjust the probability of discovering a stop event in a time interval to customize the most likely pickup hours (e.g., larger than 50%). The users also can adjust the repeat counts to customize the highly suspicious bus licenses (e.g., more than 10 repeats per month).

$$P(k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (8)$$

where $k$ is the number of occurrences, and $\lambda$ is the average rate of occurrence in a time interval.

**V. DETECTION RESULTS**

This study uses a case study to demonstrate the effectiveness of the proposed approach to detecting illegal pickup locations of intercity buses. The study area is Shenzhen City, which locates in southern China with the highest urban population density. Like many major cities in China, suppressing illegal pickups of intercity buses is of importance for Shenzhen local traffic department. The data source includes six months of GPS traces from around 2000 intercity buses managed by the local traffic department. Data is automatically collected from each bus once a few seconds.

As discussed in the last section, the illegal pickup locations will change after the illegal business grows to a certain market size and these locations are discovered by traffic police. Based
on the experience of local traffic police, we choose one month as the basic temporal unit to investigate the evolving pattern. According to the local situation and experimental results, we set up the proposed parameters shown in Table 1. Specifically, based on the observation of local traffic police, the pickup time usually is very quick (e.g., several minutes), so we set the stop duration threshold $T_{\text{stop}}$ as 30 minutes. The average length of an intercity bus is around 12 meters, thus we set grid spatial resolution $L_{\text{grid}}$ as 10 meters and bandwidth $h$ as 30 meters that is around the total length of two buses to approximate the influence range of a stop event. We set the stop density threshold $\text{DEN}_{\text{stop}}$ as 0.005 per square meters, which is equivalent to around 5 stop events in a grid based on the calculation results of the kernel density estimation with the parameters of $L_{\text{grid}}$ and $h$. The maximum distance of a waiting queue to a traffic signal determines the distance threshold to traffic signals $D_{\text{signal}}$. Local traffic department reported the speed threshold of traffic jams $V_{\text{traffic}}$. In application, these parameters should be adjusted according to different study areas.

In average, there are 300 thousands of stop events per month. As shown in Table 2, from the huge number of stop events, our approach discovers an average of 198 suspiciously illegal pickup locations per month (see the row “Total Suspicious Locations”). The average ratio among high, secondary, and low suspicious levels is 5.5:14:3:1 (see the rows “Highly Suspicious”, “Secondarily Suspicious”, and “Low Suspicious”). There are around 50% repeated suspicious locations from the last month (see the rows “Repeated Locations” and “Repeated Percentage (%))”, which validates and quantifies the evolving pattern of illegal pickup locations.

### A. Characteristics of Suspicious Pickup Locations

As shown in Figure 3a, 73% of the suspicious pickup locations have an average stop duration less than 5 minutes, which validates the quick pickup time from the site observation by traffic police. Figure 3b shows the most frequently appeared facility near a suspicious pickup location is transit bus station, followed by metro station, gas station, parking lot, and locations of travel agents. Figure 3c illustrates two peaks in the distribution of most likely pickup hours: 0:00-12:00 and 18:00-20:00. This discovered knowledge can help traffic police to understand the illegal pickup behavior.

![Characteristics of suspicious pickup locations](image)

### TABLE I. PARAMETER SETTINGS IN THE CASE STUDY

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{\text{stop}}$</td>
<td>Stop duration threshold</td>
<td>30 min</td>
</tr>
<tr>
<td>$L_{\text{grid}}$</td>
<td>Grid spatial resolution</td>
<td>10 m</td>
</tr>
<tr>
<td>$h$</td>
<td>Bandwidth</td>
<td>30 m</td>
</tr>
<tr>
<td>$\text{DEN}_{\text{stop}}$</td>
<td>Stop density threshold</td>
<td>0.005/m²</td>
</tr>
<tr>
<td>$D_{\text{signal}}$</td>
<td>Distance threshold to traffic signals</td>
<td>30 m</td>
</tr>
<tr>
<td>$V_{\text{traffic}}$</td>
<td>Speed threshold to traffic jams</td>
<td>10 km/h</td>
</tr>
</tbody>
</table>

### TABLE II. THE DETECTION RESULTS OF SUSPICIOUS PICKUP LOCATIONS

<table>
<thead>
<tr>
<th>Month</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Suspicious Locations</td>
<td>220</td>
<td>217</td>
<td>202</td>
<td>187</td>
<td>207</td>
<td>155</td>
</tr>
<tr>
<td>Highly Suspicious</td>
<td>63</td>
<td>52</td>
<td>51</td>
<td>51</td>
<td>56</td>
<td>42</td>
</tr>
<tr>
<td>Secondarily Suspicious</td>
<td>147</td>
<td>159</td>
<td>139</td>
<td>127</td>
<td>143</td>
<td>101</td>
</tr>
<tr>
<td>Low Suspicious</td>
<td>10</td>
<td>6</td>
<td>12</td>
<td>9</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>Repeated Locations</td>
<td>121</td>
<td>113</td>
<td>101</td>
<td>91</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>Repeated Percentage (%)</td>
<td>48</td>
<td>44</td>
<td>54</td>
<td>56</td>
<td>56</td>
<td></td>
</tr>
</tbody>
</table>
B. Examples of Suspicious Pickup Locations
We use two examples of suspicious pickup locations to show the different characteristics of individual cases.

The first case locates near a metro station (Figure 4a) with monthly stop events of 130. Its most likely pick hour is 9:00-10:00, with a probability of discovering a stop event up to 50% (Figure 4c). It also has three highly suspicious bus licenses with more than 10 repeated stop events per month.

The second case locates outside an entrance of a public facility (Figure 4b), with a monthly stop events of 95. Different from the first case, its most likely pickup hour is at evening, especially between 21:00-22:00, with a probability of discovering a stop event up to 33% (Figure 4c). Another difference with the first case is that there is no bus with highly repeated stop events at this location.

These two examples are both confirmed as illegal pickup locations by the local traffic police. The unique characteristics of each suspicious pickup location serve as useful knowledge for traffic police.

C. Site Investigation at Suspicious Pickup Locations
We randomly selected eight suspicious pickup locations from our detection results of February, 2013, and conducted a site investigation in March, 2013. This site investigation includes one hour of observation during most likely pickup hours. An observation of one pickup event will confirm the detection result. As Table 3 summaries, 6 out of 8 suspicious locations are confirmed by only one-hour of site observation, which demonstrates the effectiveness of the proposed detection approach.

<table>
<thead>
<tr>
<th>ID</th>
<th>Landmark</th>
<th>Investigation Time</th>
<th>Suspicious Level</th>
<th>Conclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Parking lot</td>
<td>11:00-12:00</td>
<td>high</td>
<td>unconfirmed</td>
</tr>
<tr>
<td>2</td>
<td>Gas station</td>
<td>13:00-14:00</td>
<td>high</td>
<td>unconfirmed</td>
</tr>
<tr>
<td>3</td>
<td>Overpass</td>
<td>13:00-14:00</td>
<td>middle</td>
<td>confirmed</td>
</tr>
<tr>
<td>4</td>
<td>Overpass</td>
<td>12:00-13:00</td>
<td>middle</td>
<td>confirmed</td>
</tr>
<tr>
<td>5</td>
<td>Transit bus station</td>
<td>14:00-15:00</td>
<td>high</td>
<td>confirmed</td>
</tr>
<tr>
<td>6</td>
<td>Parking lot</td>
<td>12:00-13:00</td>
<td>high</td>
<td>confirmed</td>
</tr>
<tr>
<td>7</td>
<td>Shopping mall</td>
<td>11:00-12:00</td>
<td>middle</td>
<td>confirmed</td>
</tr>
<tr>
<td>8</td>
<td>Residential area</td>
<td>14:00-15:00</td>
<td>middle</td>
<td>confirmed</td>
</tr>
</tbody>
</table>

VI. WEBSITE-BASED USER INTERFACE

To help traffic police better understand the detection results, the web-based user interface of the GIS-TP accommodates several major functionalities. As demonstrated in Figure 5, with a selected year and month, it can display suspicious pickup locations on a map with the suspicious levels indicated by different colors. It allows users to click a suspicious pickup location to check its information such as suspicious level, total stop events, average stop duration, most likely pickup hours, and highly suspicious bus licenses. Users can also select a bus license to observe its whole GPS trace on the map.

Figure 5. The web-based user interface

VII. CONCLUSIONS AND FUTURE WORK

This study presents an initial approach for assisting in suppressing illegal pickups of intercity buses with the analysis of GPS traces. We implement this approach in a system called GIS-TP. This system integrates the computation power of cloud computing platform, the efficiency of GIS engine for spatial analysis, and the intuitive visualization & interaction interface of web-based GIS service, together transforming massive GPS data to actionable intelligence. The case study in Shenzhen City has demonstrated the effectiveness of the GIS-TP with its high accuracy of detecting illegal pickup locations, and its functionality to help traffic police to understand illegal pickup behavior and decision-making for a site investigation. The local traffic police highly appreciated this system, and have already been using it in practice.

There are a few possible extensions to this initial effort. Firstly, the traffic conditions inferred from floating car data (FCD, e.g., taxi GPS data) can help to improve the accuracy of recognizing stops caused by traffic jams. Secondly, more intensive site investigations could help formulate more characteristics of illegal pickups, and offer a larger training dataset. With a larger training dataset, we can develop machine learning models to further improve the detection accuracy. Thirdly, for buses equipped with a door inspection system, we can incorporate the signal of a door-open event to infer whether a stop event is related with passenger pickups. At last, although the proposed system uses historical GPS data as the input, it can be integrated with real-time vehicle tracking system. With the real-time GPS stream, it can compare current stops with the suspicious ones in the knowledge database, and make real-time alerts to monitor officers, so they can immediately take actions such as checking the bus video, warning a suspicious bus, and etc.

ACKNOWLEDGMENT

The authors thank Jay Lee and Xia Huang for their useful suggestions on this article.
REFERENCES

[2] Idiagnosis, Available at: solutions/intercitybuseservice.html

2167