Urban Transportation Analysis
Using Taxi Trajectory and Weather Data

A dissertation submitted
to Kent State University in partial
fulfillment of the requirements for the
degree of Master of Digital Sciences

by

Hui Tang

October 2016
Thesis written by

Hui Tang

B.S., Peking University, China 1996
Ph.D., Stony Brook University, 2007
M.D.S., Kent State University, 2016

Approved by

______________________________, Cheng-Chang Lu, Advisor
______________________________, Chair, School of Digital Sciences
Abstract

Urban data has been collected and accumulated. It greatly advances our knowledge and understanding of urban transportation patterns and human mobility behaviors. In this thesis, both the taxi trajectory and weather data for transportation features of Hangzhou, a city in China, were analyzed. First, a study of characteristics of traffic attributes, such as traffic speed, and records of taxi pickups/drop-offs, was carried out by analyzing their temporal changes on selected individual streets and the whole city. Second, an analysis among these attributes was performed to investigate whether or not there are correlation behavior among them. Finally, a supervised machine learning model, support vector machine (SVM), was utilized to make a prediction on the traffic situations in selected streets from multiple types of transportation and weather attributes. Data analysis included in this thesis was performed over massive taxi trajectory and weather data with a set of data analytics tools in the environment of R package. The results show that data statistics, visualization, and machine learning tools can be applied to a variety types of urban data, facilitating domain users and city residents to gain knowledge of a city’s transportation features.
# TABLE OF CONTENTS

LIST OF FIGURES ........................................................................................................ VI

LIST OF TABLES .......................................................................................................... VII

CHAPTER 1 ......................................................................................................................... 1

1.1 Data Collection and Preparation ............................................................................. 2

1.2 Data Analysis Tasks ................................................................................................. 2

CHAPTER 2 RELATED WORK ......................................................................................... 4

CHAPTER 3 URBAN DATA COLLECTION AND PREPARATION .............................. 6

3.1 Taxi Trajectory Data ................................................................................................. 6

3.2 Weather Data .......................................................................................................... 8

CHAPTER 4 URBAN TRANSPORTATION ANALYSIS ............................................ 10

4.1 Data Analysis Tool ................................................................................................. 10

4.2 Data Characteristics Study ....................................................................................... 10

4.2.1 Daily Traffic Speed Pattern ............................................................................... 11

4.2.2 Hourly Traffic Speed Pattern ........................................................................... 13

4.2.3 Daily Taxi Service Pattern .................................................................................. 15

4.2.4 Hourly Taxi Service Patterns ............................................................................. 15

4.2.5 Location-based Patterns ...................................................................................... 17

4.2.6 Weather Patterns ............................................................................................... 20

4.3 Data Relationship Study ......................................................................................... 21

4.3.1 Location-based Correlation of Traffic Speed ....................................................... 22

4.3.2 Location-based Correlation of Traffic Speed and Taxi Service ............................ 25
4.3.3 Influence of Weather to Transportation................................................................. 26
4.4 SVM Model Training and Prediction .......................................................................... 27
4.4.1 Support Vector Machine......................................................................................... 27
4.4.2 SVM Prediction of Street Traffic Speed................................................................. 28

REFERENCES......................................................................................................................... 35
LIST OF FIGURES

Figure 1 Daily average traffic speed................................................................. 11
Figure 2. Variation of the daily average traffic speed. ................................. 12
Figure 3. R code of daily average speed analysis........................................ 12
Figure 4. Daily average traffic speed............................................................... 13
Figure 5. Hourly average speed................................................................. 14
Figure 6. R code of hourly average speed analysis.................................... 14
Figure 7. Daily taxi pickups......................................................................... 15
Figure 8. Average hourly taxi pickups........................................................... 16
Figure 9. Average amount of actively running taxis ................................ 17
Figure 10. Six selected locations on the map of Hangzhou............................ 18
Figure 11. Daily average speed...................................................................... 19
Figure 12. Average speed over different hours of six different locations ........ 20
Figure 13. Weather summary from hourly data............................................ 21
Figure 14. Daily temperature....................................................................... 21
Figure 15. Heat map view of the correlation matrix................................ 23
Figure 16. R implementation code of location-based correlation..................... 24
Figure 17. Correlation between traffic speed and taxi drop-offs.................... 25
Figure 18. Linear Classifier.......................................................................... 28
Figure 19. Plot of SVM results...................................................................... 32
LIST OF TABLES

Table 1. Computing a set of transportation attributes over each street segment. .......... 8
Table 2. Average attributes computed on all streets of Hangzhou .......................... 8
Table 3. Hourly weather information ...................................................................... 9
Table 4. Selected locations in Hangzhou ................................................................. 18
Table 5. Correlation matrix of the six locations for their daily traffic speed ............... 23
Table 6. Sample Traffic Data Used for SVM .......................................................... 29
Table 7. Matching of Predicted and Real Pickups in Four Classes ............................ 34
Chapter 1

Introduction

The rapid advancement of sensing technologies and computing infrastructures greatly help researchers to collect massive trajectory data of people and vehicles in urban spaces at an unprecedented scale and speed. With the prevalent GPS, Wi-Fi, Cellular, and RFID devices, human mobility information is instantly recorded as the moving paths of taxis, fleets, public transits, and mobile phones. The data is utilized to analyze urban system, environment, and economy to improve urban planning and operations.

The taxi trajectory data is of great interest for the observation, evaluation, and optimization of transportation infrastructures and policies. For example, major problems in modern cities, such as traffic jams, unbalanced capacities, frequently occurring accidents, and unsatisfied transportation services, are caused by improper road planning, maintenance, and traffic control. Researchers and analysts need to evaluate such situations in transportation studies. The emerging taxi trajectory data provides useful information from which the statistics of real traffic flow can be extracted and city-wide transport patterns can be discovered. Urban computing has emerged recently in the data mining community to advance knowledge discovery from a variety of data including taxi trajectories [1].

Weather is an important factor that influences the traffic situation and transportation performance. Researchers have investigated the relation of weather information and traffic events (such as crash occurrence) [2]. However, the weather data has not been well connected
with the taxi trajectory data to analyze and predict the dynamics of urban transportation. In this thesis, we perform a set of data analysis tasks by integrating massive taxi trajectory data with hourly weather data to investigate traffic status and taxi services.

1.1 Data Collection and Preparation

We utilize the real world taxi data from Hangzhou, a city in China. A large set of taxi trajectories and related weather information are collected and prepared to meet the need of data analysis. The taxi dataset is sampled in three months by 8,120 taxis in Hangzhou. It has a raw size of around 230GB. The geometric data of Hangzhou’s road network consists of 14,639 street segments. The raw data is preprocessed and map-matched onto the street segments. A series of attributes of traffic and attributes of taxi are computed for each street segment. The value of these attributes changes every two hours. The hourly weather data is obtained from an online service of historical weather information, including a variety of meteorological attributes of Hangzhou recorded in a city center location.

1.2 Data Analysis Tasks

The taxi trajectory and weather data were analyzed regarding transportation features of Hangzhou. The transportation analysis consists of three tasks. First, a study was carried out to analyze the data characteristics of traffic information, including the average traffic speed on individual streets of interests and average traffic speed throughout the city. Taxi service information, passenger pick-up and drop-off information, was also analyzed to show the mobility pattern of city residents. The above information varies over time, reflecting the urban dynamics of different days and different periods of a day.
Second, we perform data correlation and regression analysis among multiple attributes of traffic information, taxi service information, and weather information. We conduct a set of tests to investigate whether or not the attributes have correlative behaviors such as the weather type and traffic speed.

Third, we utilize the supervised machine learning model, SVM (support vector machine), to analyze and predict the transportation situations. We perform a series of tests to study whether or not the traffic, taxi, and weather attributes in some places can be used to predict the situation in other places.
Chapter 2

Related Work

Exploring patterns and trends of intra-urban human mobility advances the understanding of urban dynamics and reveals socioeconomic driving forces [3, 4, 5]. The increasing availability of GPS data has greatly facilitated the study of street networks [6, 7]. Urban traffic flow can be viewed as transportation demands which are aggregately distributed in street networks [8]. Taxis often serve as floating cars to obtain human mobility data and examine real-time traffic status and individual behaviors [9, 10, 11, 12].

In the field of data mining, trajectories of human and vehicle motions are used to discover knowledge from large-scale datasets [1]. Utilizing the trajectory data has been divided into three main categories: the study of the collective behavior of a city's population, the traffic flow, and the operators (e.g. drivers) [13]. In particular, vehicle trajectory data has been used in traffic monitoring and prediction [14], urban planning [15], driving routing [16, 17], extracting geographical borders [18], service improvement [19], energy consumption analysis [20], and dynamic travel time estimation [21]. Large-scale mobile phone data with GIS information is used to uncover hidden patterns in urban road usage [22], find privacy bounds of human mobility [23], estimate travel time [24] and infer land use [25]. Public transit trajectories are used in bus arrival time predictions [26], and travelers' spending optimization [27]. NEAT [28] studies trajectory clustering over road networks by considering traffic flows together with segment densities and connectivity.
A large number of approaches have been proposed to visually explore movement data (see [29] for a recent survey). Many of them are focused on the origins and destinations of the trajectories, such as flow maps [30], Flowstrates [31], OD maps [32], and visual queries for origin and destination data [33]. Other work visualizes trajectories using various visual metaphors and interactions, such as GeoTime [34], TripVista [35], FromDaDy [36], vessel movement [37], route diversity [38], and more [39, 40, 41, 42, 43, 44, 33]. Some of these approaches coordinate multidimensional visualization and map views [35, 33, 44].

Weather conditions can greatly affect the traffic patterns, transportation efficiency, and safety. Hourly rainfall data is used to represent and indicate daily pavement surface status [45, 46]. For traffic accidents, the number of rainy days and the amount of rain are used to identify its relation with crashes [47, 48, 2]. In this thesis, we utilize both taxi trajectory and weather data to analyze urban transportation features.
Chapter 3

Urban Data Collection and Preparation

3.1 Taxi Trajectory Data

A massive amount of trajectory data sets is available today. Some of them are available for public use, such as the Beijing city taxi data [49], the Rome Italy taxi data [50], and the New York City taxi data [51]. While many trajectory datasets are not publicized, including taxis [52], public transits [53], human paths [54], and many more. Researchers may use those data sets with permission for their studies and publish their discovery based on the data. With the widespread application of trajectory recording technologies, these datasets will be used in many fields for the purpose of research.

The trajectory data used in this thesis comprises daily trajectories of 8,120 taxis in Hangzhou. Hangzhou is a big city in China with a population over 9 million residents in a condensed area, where taxi is one of the major traffic modes of resident. Nearly three thousand GPS sample positions are reported by one taxi every day. Each sample consists of taxi plate, time, status, speed, direction, and latitude and longitude, e.g., (AXXXXX, 12/27/2011 13:19:32, 0, 50, 180, 30.289, 120.168). The data used in this study spans three months, December 2011, January 2012, and February 2012. The accumulated taxi trajectories create a data set with a raw size of 230GB.

The corresponding road network contains 14,639 street segments. The long street is segmented into smaller ones for accurate spatial representation. Raw trajectory data is structured
by computing a set of transportation attributes over street segments. As shown in Table 1, a
unique identifier was assigned to each street segment. For each street, there are information of
street name, street length, and the geolocation of its center (center_latitude and
center_longitude). For a given day and a given time period (i.e., two hours of the day starting
from the value hour), the traffic attributes are computed to obtain the average traffic speed over
the time period and the average travel time. Meanwhile, the amount of taxi pickups (PN) and
drop offs (DN), and the amount of taxis passing the street (flow) were calculated based on the
taxi attributes. The computation was implemented by mapping each GPS point onto the closest
street segment, followed by accumulating these attributes in the given time window over each
segment. For example, locations of pickups/drop offs were extracted from taxi trajectory data.
For each period, T1 to Tn, counts of such activities over street segments were computed. T1 to
Tn was set as two hours. Following the similar way, average travel speed over street segments
was also computed over a period of T1 to Tn.

Data needs to be extracted and stored in a way that facilitating specific goal of data
analytics. I wrote programs using Python or R to organize the data and the resulting datasets
were stored as tables. Table 2 shows an example of the average values computed over all street
segments of Hangzhou.
Table 1. Computing a set of transportation attributes over each street segment.

<table>
<thead>
<tr>
<th>id</th>
<th>center_lat</th>
<th>center_long</th>
<th>length</th>
<th>DN</th>
<th>PN</th>
<th>flow</th>
<th>speed</th>
<th>travelTime</th>
<th>name</th>
<th>day</th>
<th>hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>6165</td>
<td>30.23199</td>
<td>120.2101</td>
<td>1.389167</td>
<td>1</td>
<td>1</td>
<td>206</td>
<td>65.62</td>
<td>0.021169</td>
<td>XixingBridge</td>
<td>12/10/2011</td>
<td>0</td>
</tr>
<tr>
<td>14792</td>
<td>30.2702</td>
<td>120.1664</td>
<td>0.913979</td>
<td>2</td>
<td>8</td>
<td>323</td>
<td>58.21</td>
<td>0.0157</td>
<td>ShangtanGaojialu</td>
<td>12/10/2011</td>
<td>4</td>
</tr>
<tr>
<td>15082</td>
<td>30.25925</td>
<td>120.1738</td>
<td>0.285974</td>
<td>3</td>
<td>0</td>
<td>17</td>
<td>14.59</td>
<td>0.019603</td>
<td>Diyihospital</td>
<td>12/10/2011</td>
<td>16</td>
</tr>
</tbody>
</table>

Table 2. Average attributes computed on all streets of Hangzhou

<table>
<thead>
<tr>
<th>day</th>
<th>hour</th>
<th>AverageDN</th>
<th>Averagespeed</th>
<th>Averageflow</th>
</tr>
</thead>
<tbody>
<tr>
<td>12/10/2011</td>
<td>10</td>
<td>1.512636612</td>
<td>31.267</td>
<td>26.65219</td>
</tr>
<tr>
<td>12/10/2011</td>
<td>12</td>
<td>1.354098361</td>
<td>31.47046</td>
<td>26.15587</td>
</tr>
<tr>
<td>12/10/2011</td>
<td>14</td>
<td>1.276502732</td>
<td>30.99558</td>
<td>25.6403</td>
</tr>
<tr>
<td>12/10/2011</td>
<td>16</td>
<td>1.099863388</td>
<td>30.46372</td>
<td>24.9362</td>
</tr>
<tr>
<td>12/10/2011</td>
<td>18</td>
<td>1.688729508</td>
<td>31.594</td>
<td>27.93921</td>
</tr>
</tbody>
</table>

3.2 Weather Data

The weather data was obtained via an online service (Forecase.io) of historical weather information. The Forecast API allows users to look up the weather all over the world. A historical data request yields the observed weather at a given time in past years. The returned information is organized as data points. Each data point object contains various properties, representing a particular weather phenomenon occurring at a specific point in time. The weather attributes retrieved from this on-line service include pressure, visibility, humidity, temperature, dewPoint, windSpeed and summary (cloudy, rain, etc.). Table 3 shows an example of the weather data in a particular day.
Table 3. Hourly weather information

<table>
<thead>
<tr>
<th>pressure</th>
<th>visibility</th>
<th>humidity</th>
<th>temperature</th>
<th>dewPoint</th>
<th>summary</th>
<th>windSpeed</th>
<th>day</th>
<th>hour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1029.37</td>
<td>7.24</td>
<td>0.82</td>
<td>5.22</td>
<td>2.43</td>
<td>Mostly Cloudy</td>
<td>4.23</td>
<td>12/1/2011</td>
<td>7</td>
</tr>
<tr>
<td>1029.37</td>
<td>8.1</td>
<td>0.82</td>
<td>5.22</td>
<td>2.34</td>
<td>Mostly Cloudy</td>
<td>4.21</td>
<td>12/1/2011</td>
<td>8</td>
</tr>
<tr>
<td>1030.36</td>
<td>7.84</td>
<td>0.77</td>
<td>6.1</td>
<td>2.31</td>
<td>Mostly Cloudy</td>
<td>4.98</td>
<td>12/1/2011</td>
<td>9</td>
</tr>
<tr>
<td>1030.36</td>
<td>8.1</td>
<td>0.76</td>
<td>6.26</td>
<td>2.34</td>
<td>Mostly Cloudy</td>
<td>5.14</td>
<td>12/1/2011</td>
<td>10</td>
</tr>
<tr>
<td>1030.36</td>
<td>7.21</td>
<td>0.76</td>
<td>6.3</td>
<td>2.34</td>
<td>Mostly Cloudy</td>
<td>5.94</td>
<td>12/1/2011</td>
<td>11</td>
</tr>
</tbody>
</table>
Chapter 4

Urban Transportation Analysis

4.1 Data Analysis Tool

The data analysis tasks were carried out in a R environment. This tool provides a wide variety of statistical functions with high efficiency. In addition, R has powerful graphics facilities for the production of high quality diagrams. Besides the statistic computing functions, three extension libraries, Sqldf, ggplot2, and e1071, were utilized in the study included in this thesis.

Sqldf package provides a mechanism to manipulate R data frames using SQL, allowing users to use abundant SQL statements to perform tasks over data tables in the R environment. ggplot2 package is a widely used in plotting system, offering abundant base and lattice graphics to visualize results of data analysis. It has a powerful model of graphics that makes it easy to produce complex multi-layered graphics with easy interface and seamless integration with data frames. e1071 is a package in R that has specialized functions for implementing Naïve Bayes (conditional probability), SVM, Fourier Transforms, Bagged Clustering, Fuzzy Clustering, etc. In this thesis work, it was used for SVM training and prediction.

4.2 Data Characteristics Study

The re-structured data, representing transportation characteristics of Hangzhou, was analyzed to investigate traffic and human mobility patterns.
4.2.1 Daily Traffic Speed Pattern

The average traffic speed of the city in one given day was obtained by calculating the mean of average speeds of every two hours in that day. This speed reflects the global traffic information of Hangzhou. Figure 1 shows the daily traffic speed of Hangzhou during December 2011 in a bar chart, suggesting that there is no difference in daily traffic speed in December 2011 in the city.

![Bar chart showing daily traffic speed in December 2011](image)

**Figure 1 Daily average traffic speed (km/hour) of Hangzhou in Dec. 2011.**

Next, I want to know whether there is a pattern in the temporal change of the citywide daily traffic speed. As shown in figure 2, a different plotting method was applied to the dataset to show the variation in daily traffic speed. A pattern of speed change over the month was observed, four peaks corresponding to weekends of Dec 2011. This pattern clearly indicates that traffic speed is higher during weekends than weekdays. The implementation code in R of the analysis is shown in Figure 3.
Figure 2. Variation of the daily average traffic speed (km/hour) of the whole city in Dec 2011.

```r
w<-read.csv("AverageInfo_Decspeed.csv", header = TRUE, sep = ",")
out <- sqldf("select day, avg(Averagespeed) as av from w group by day")
ggplot(data=out, aes(x=day, y=av, group=1))+geom_line(colour="blue", size=1)+ylab("Average Speed")+xlab("Date")
+theme(axis.title=element_text(size=16,face="bold"),axis.text.x=element_text(face="bold",color="black",size=14,angle=45),axis.text.y=element_text(face="bold", color="black",size=14, angle=0))
```

Figure 3. R code of daily average speed analysis

Figure 4 shows the daily traffic speed in January and February of 2012. Compared with the pattern observed in December, 2011, a peak with a longer period across the intersection of the two months showed up. It is due to the holiday season, Chinese New Year. This holiday season changes the mobility patterns of Chinese people for up to one month before and after the Chinese
New Year’s day on January 23, 2012. During the holiday season, the traffic speed became faster than usual because many residents leave big cities where they work and go back to their hometowns for family reunion. Therefore, there are less vehicles and smaller population appearing on the streets than usual. The weekly pattern recovered around Feb 11, Feb 18 and Feb 25 after the busy holiday for city residents.

![Graph](image)

Figure 4. Daily average traffic speed (km/hour) of Hangzhou in Jan and Feb 2012

4.2.2 Hourly Traffic Speed Pattern

To investigate 24-hour fluctuation of traffic speed in Hangzhou, hourly average speed of all the street segments was calculated. Figure 5 shows the variation in hourly average speed in three months: December 2011 (red), January 2012 (blue), and February 2012 (green). Trend in the variation of hourly traffic speed from the above three months follows the same pattern: average hourly speed is higher at night than during the day, valley points present at around 8am and 4pm,
indicating that two lowest traffic speeds occur at rush hours. The implementation R code is shown in Figure 6.

**Figure 5.** Hourly average speed (km/hour) in Hangzhou in Dec, 2011 (red), Jan, 2012 (blue), and Feb, 2012 (green).

```r
# Load data
d <- read.csv("AverageInfo_Decspeed.csv", header = TRUE, sep = ",")

dout <- sqldf("select hour, avg(Averagespeed) as av from d group by hour")

j <- read.csv("AverageInfo_Janspeed.csv", header = TRUE, sep = ",")

jout <- sqldf("select hour, avg(Averagespeed) as av from j group by hour")

f <- read.csv("AverageInfo_Febspeed.csv", header = TRUE, sep = ",")

fout <- sqldf("select hour, avg(Averagespeed) as av from f group by hour")

# Plot
ggplot(data=d, aes(x=hour, y=av, group=1))+geom_line(colour="Red", size=1)+geom_line(data = jout, colour="Blue", size=1)+geom_line(data = fout, colour="Green", size=1)+ylab("AverageSpeed") +xlab("Hour")+theme(axis.title=element_text(size=16,face="bold"),axis.text.x=element_text
(face="bold",color="black",size=14, angle=0),axis.text.y=element_text(face="bold", color="black", size=14, angle=0))
```

**Figure 6.** R code of hourly average speed analysis
4.2.3 Daily Taxi Service Pattern

Similar to traffic speed, the daily taxi service attributes over the whole city were computed. The pattern of taxi service was determined by analyzing the taxi pickups or drop offs on each street. Figure 7 is the plot of daily taxi pickups of Hangzhou in January, 2012. The curve hits the valley point on January 22, 2012, a day before Chinese New Year’s day, and a steady increase in taxi pickups starts on January 23, 2012. The data indicates that the minimum need of taxi service took place on the day before Chinese New year’s day.

![Figure 7. Daily taxi pickups in Jan, 2012](image)

4.2.4 Hourly Taxi Service Patterns

Variation of taxi service in a day was investigated by analyzing taxi pickups/drop offs. The total hourly taxi pickups/drop offs for each day were computed. This value was averaged over one month. The data points of average hourly pickups in December of 2011 (red), January of 2012 (blue), and February of 2012 (green) are shown in figure 8. The pattern of taxi service is
that the peak pickups occurs at 8:00 pm, the valley pickups is at 4:00 am, and pickups are steady during the day time. Figure 9 further shows the average amount of actively running taxis (i.e., taxi flow) over each hour from one month’s taxi data. This figure shows the similar trend of the taxi pickups, indicating that taxi drivers fit in the human mobility pattern.

Figure 8. Average hourly taxi pickups, Dec, 2011 (red), Jan, 2012 (blue) and Feb, 2012 (green)
4.2.5 Location-based Patterns

To find the location-based patterns of taxi service, a study of the relationship of traffic speed and taxi service was carried out using trajectory data collected from selected locations in Hangzhou. The location based analysis is useful for city residents and administration to understand the transportation attributes over hotspots, landmarks, and important streets/bridges, etc. This work was conducted by: (1) finding street segments closest to the selected locations; (2) extracting the transportation attributes for these streets from the data tables; and (3) analyzing the attributes with the analytical and visualization tools. The selected locations are listed in table 4, and mapped out in Figure 10.
Table 4. Selected locations in Hangzhou

<table>
<thead>
<tr>
<th>Street Segment ID</th>
<th>Street Name</th>
<th>Location Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>12150</td>
<td>Gushanlu</td>
<td>Gushan, the most popular tourist center</td>
</tr>
<tr>
<td>12500</td>
<td>Tianmushanlu</td>
<td>The administration building of Hangzhou</td>
</tr>
<tr>
<td>6165</td>
<td>XixingBridge</td>
<td>A highway bridge connecting two parts of the city</td>
</tr>
<tr>
<td>15082</td>
<td>DiyiHospital</td>
<td>A big hospital</td>
</tr>
<tr>
<td>12198</td>
<td>Zhedalu</td>
<td>Zhejiang University</td>
</tr>
<tr>
<td>14792</td>
<td>Shangtanggaojialu</td>
<td>A busy section of an expressway of the city</td>
</tr>
</tbody>
</table>

Figure 10. Six selected locations on the map of Hangzhou
The daily average speed of the six locations in January, 2012 is shown in figure 11. It was obtained using traffic speed of taxis on those roads. The speed of Tianmushanlu (red) is mostly close to zero. This is because that taxi is not allowed to pass that road due to a traffic restriction on taxis. A stable high speed was observed on XixingBridge (Green), which is a highway, connecting two parts of the city. In contrast, the Shangtanggaojialu, another highway, a burst in speed occurred before and after Jan 23, the Chinese New Year’s day. These patterns suggests that XixingBridge and Shangtanggaojialu play different role in the transportation of Hangzhou.

Figure 12 further depicts the hourly change of the average speed over the six locations. The curves generally follow the typical pattern for traffic during the day and at night. Although Shangtanggaojialu (black line) is a highway, a low traffic speed, 40 km/hour, was observed during the day. Traffic jams frequently appear in this section.

![Figure 11. Daily average speed (km/hour) of six different locations at Hangzhou. Blue: Gushanlu Red: Tianmushanlu Green: XixingBridge Purple: DiyiHospital Yellow: Zhedalu Black: Shangtanggaojialu](image-url)
Figure 12. Average speed (km/hour) over different hours of six different locations at Hangzhou. Blue: Gushanlu Red: Tianmushanlu Green: XixingBridge Purple: DiyiHospital Yellow: Zhedalu Black: Shangtanggaojialu

4.2.6 Weather Patterns

Figure 13 shows the percentages of different weather summary types during the three months mentioned earlier. An example of temperature change of January, 2012 is shown in figure 14. Climate change in January of 2012 is not dramatic in temperature and type of weather. Most of the days are either foggy or cloudy, with steady low temperature.
Figure 13. Weather summary from hourly data during the period.

Figure 14. Daily temperature in Jan, 2012

4.3 Data Relationship Study

The analysis of the correlation is an important task to find relationship between different data sets. Weather plays important role in urban public transportation. In this thesis, I want to know the relationship between taxi service and weather condition.
4.3.1 Location-based Correlation of Traffic Speed

The correlation analysis was first performed on daily average speed (Jan 2012) of the six selected locations. This task is conducted using Pearson correlation. Pearson correlation measures the linear dependence between two variables, giving a value between +1 and -1. +1 represents total positive linear correlation, 0 is no linear correlation, and -1 is total negative correlation. Pearson’s correlation coefficient of two variables, X and Y, is calculated as shown in equation 1, where $X$ and $Y$ are the means of $X$ and $Y$, respectively.

$$
    cor(X,Y) = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{(X - \bar{X})(X - \bar{X}) + (Y - \bar{Y})(Y - \bar{Y})}}
$$

**Equation 1. Pearson correlation of two variables**

For the given locations, Pearson correlation coefficient was computed using data sets from a pair of the six locations. Then a correlation matrix is shown in Table 5. The matrix was further visualized using a heat map, as shown in Figure 15. Here, I gained useful knowledge to understand the relationship between different locations. For example, Gushanlu, a tourism site, generally shows negative correlation to other locations. Zhedalu are positively correlated to Shangtanggaojialu, with a high Pearson correlation coefficient of 0.87. The implementation code in R is shown in Figure 16.
Table 5. Correlation matrix of the six locations for their daily traffic speed

<table>
<thead>
<tr>
<th></th>
<th>Gushanlu</th>
<th>Tianmushanlu</th>
<th>XixingBridge</th>
<th>DiyiHospital</th>
<th>Zhedalu</th>
<th>Shangtanggaojiaju</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gushanlu</td>
<td>1</td>
<td>-0.25</td>
<td>-0.44</td>
<td>-0.28</td>
<td>-0.57</td>
<td>-0.54</td>
</tr>
<tr>
<td>Tianmushanlu</td>
<td>-0.25</td>
<td>1</td>
<td>0.19</td>
<td>0.2</td>
<td>0.05</td>
<td>-0.1</td>
</tr>
<tr>
<td>XixingBridge</td>
<td>-0.44</td>
<td>0.19</td>
<td>1</td>
<td>0.2</td>
<td>0.49</td>
<td>0.6</td>
</tr>
<tr>
<td>DiyiHospital</td>
<td>-0.28</td>
<td>0.2</td>
<td>0.2</td>
<td>1</td>
<td>0.37</td>
<td>0.54</td>
</tr>
<tr>
<td>Zhedalu</td>
<td>-0.57</td>
<td>0.05</td>
<td>0.49</td>
<td>0.37</td>
<td>1</td>
<td>0.87</td>
</tr>
<tr>
<td>Shangtanggaojiaju</td>
<td>-0.54</td>
<td>-0.1</td>
<td>0.6</td>
<td>0.54</td>
<td>0.87</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 15. Heat map view of the correlation matrix in Table 5.
Figure 16. R implementation code of location-based correlation.
4.3.2 Location-based Correlation of Traffic Speed and Taxi Service

Correlation analysis was performed on traffic speed and taxi services to determine whether traffic speed has impact on taxi pickups/drop offs. Pearson correlation coefficient was computed using data of average traffic speed and average pickup/drop-offs over the whole city. The result shows that there is no significant correlation between them. The correlation analysis was then performed on data from individual selected streets. Figure 17 shows the results of the computation on the dataset of the January, 2012. The location based correlations vary from street to street. There are relatively strong negative correlations among Gushanlu, Zhedalu and Shangtanggaojialu, indicating that as the traffic speed increase, less taxi pickups/drop offs occur. The positive correlation observed in Tianmushanlu can be neglected since a traffic restriction was applied to taxi service in this location.

![Correlation graph](image)

**Figure 17.** Correlation between traffic speed and taxi drop-offs at six important streets.
4.3.3 Influence of Weather to Transportation

It would be of great interest to find whether the weather has a significant influence on urban transportation. A set of correlation analysis was conducted using weather dataset and attributes of traffic speed and taxi service.

Pearson correlation analysis was performed on hourly temperature and hourly taxi pickups of December, 2011. Same approach was carried out on wind speed and taxi pickups. The results would answer the question: is wind or temperature a factor that affects residents choose taxis as their transportation mode? The nearly zero value of the coefficients, -0.03 and -0.005, suggests there is no significant correlation between weather and taxi pickups. In other words, residents may not alter their travel methods due to the temperature and wind speed. Alternatively, variations of temperature and wind are not big enough to affect the resident’s decision on transportation mode.

To determine whether the travel pattern of residents can be affected by weather conditions, such as sunny, cloudy, and rainy. Figure 18 shows the average pickups under different weather conditions per hour at the six selected locations. The data used is collected from 6:00 am to 8:00 pm. The value of each pie is the number of pickups. Lowest average taxi pickups showed up in foggy days, indicating that there is lower need in taxi pickup in winter when it is foggy than other weather conditions.
Figure 18. Average taxi pickups per hour over the six selected locations of different weather conditions

4.4 SVM Model Training and Prediction

4.4.1 Support Vector Machine

SVM is introduced by Boser, Guyon and Vapnik [55]. It has been applied to a wide variety of the classification functions. It has become a popular machine learning tool due to its empirically good performance. SVM has achieved many successful applications in the fields of bioinformatics, text and data mining, image recognition, etc.

Machine learning is about learning hidden patterns from data. For an input domain $X$ and an output domain $Y$, a training set is given as $(x_1, y_1), (x_2, y_2) \ldots (x_m, y_m)$. SVM wants to learn a classifier $y = f(x, p)$ where $p$ is the function parameters. For example, in Figure 18, a linear classifier is used as a separation line $y = f(x, \{k, b\}) = \text{sign}(kx + b)$. SVM try to learn $f$ by choosing a function that performs well for the classification of the training set. In general, it minimizes the
error \( E(p) = \frac{1}{m} \sum_{i=1}^{m} e(f(x_i, p), y_i) \), where \( e() \) is the loss function that measures the difference between the computed value and the training value. \( p \) is inferred from the minimization and then the classifier \( f \) is defined. In SVMs, the types of \( f \) need to be specified, such as the linear classifier or nonlinear classifiers. A key feature of nonlinear classifiers is to map data into a richer feature space including nonlinear features. That is, in SVM training, \( x \) is mapped to \( \phi(x) \) by using common kernels like radial basis functions [56]. When \( y \) is not a signed function or represents discrete classes, the SVM perform regression instead of classification [56]. Both SVM regression for continuous output value (traffic speed) and discrete classes (taxi pickups) were applied in the analysis shown in the following sections.

![Linear Classifier](image)

**Figure 18. Linear Classifier.**

### 4.4.2 SVM Prediction of Street Traffic Speed

SVM was used to train and predict location based traffic speed. The hourly average traffic speed of four streets, Shangtanggaojialu, Zhedalu, Gushanlu, and XixingBridge, were
included, together with two weather attributes, visibility and temperature. Tianmushanlu and Diyihospital were not included due to traffic restriction on taxi service. Table 6 shows an example of data used in the SVM, and the data were collected from December 2011 to February 2012. S_speed (Z_speed, G_speed, X_speed) represents the average traffic speed at Shangtanggaojialu (Zhedalu, Gushanlu, XixingBridge) for individual hour.

Table 6. Sample Traffic Data Used for SVM

<table>
<thead>
<tr>
<th>G_speed</th>
<th>S_speed</th>
<th>Z_speed</th>
<th>X_speed</th>
<th>hour</th>
<th>visibility</th>
<th>temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>29.77385</td>
<td>53.40853</td>
<td>30.18507</td>
<td>67.96977</td>
<td>6</td>
<td>2.29</td>
<td>2.19</td>
</tr>
<tr>
<td>20.12163</td>
<td>35.75136</td>
<td>22.46074</td>
<td>64.60191</td>
<td>8</td>
<td>1.85</td>
<td>2.79</td>
</tr>
<tr>
<td>16.33407</td>
<td>20.07531</td>
<td>15.27104</td>
<td>63.22178</td>
<td>10</td>
<td>2.37</td>
<td>4.99</td>
</tr>
<tr>
<td>12.37671</td>
<td>20.13178</td>
<td>23.47837</td>
<td>61.94536</td>
<td>14</td>
<td>2.75</td>
<td>7.82</td>
</tr>
<tr>
<td>14.77984</td>
<td>38.10515</td>
<td>28.84262</td>
<td>64.11703</td>
<td>16</td>
<td>2.14</td>
<td>7.78</td>
</tr>
<tr>
<td>16.69815</td>
<td>42.68706</td>
<td>30.52768</td>
<td>63.48749</td>
<td>18</td>
<td>3.01</td>
<td>6.21</td>
</tr>
<tr>
<td>20.02816</td>
<td>41.3024</td>
<td>30.16416</td>
<td>66.4867</td>
<td>20</td>
<td>2.74</td>
<td>6.07</td>
</tr>
<tr>
<td>27.02882</td>
<td>52.21089</td>
<td>33.33238</td>
<td>67.8327</td>
<td>22</td>
<td>2.69</td>
<td>3.94</td>
</tr>
</tbody>
</table>

SVM training and prediction on above datasets were achieved in six steps.

Step 1: Generate training and testing sets by partitioning the dataset

```r
#randomly select samples from the data table for training
smp_size <- floor(0.75 * nrow(d)) #75% of the hourly data is selected
# set the seed to make the data table partition reproductible
set.seed(123)
train_ind <- sample(seq_len(nrow(d)), size = smp_size)
# generate two sets for training (75%) and testing (25%)
train <- d[train_ind, ]
test <- d[-train_ind, ]
```
Step 2: Define the output value to be predicted and the input value

```R
# define x and y set from the training set
# here Z_speed is the output value for prediction
x <- subset(train, select = -Z_speed)
y <- subset(train, select = Z_speed)
```

Step 3: Run SVM training with x and y

```R
svm1 = svm(x, y)
summary(svm1) # show the SVM model
```

The `svm1` summary is shown as following.

```
svm.default(x = x, y = y)
Parameters:
  SVM-Type:  eps-regression
  SVM-Kernel:  radial
  cost:  1
  gamma:  0.1666667
  epsilon:  0.1
  Number of Support Vectors:  400
```

Step 4: Perform prediction using the test dataset

```R
# generate test x and y with Z_speed as the output value for prediction
tx <- subset(test, select = -Z_speed)
ty <- subset(test, select = Z_speed)
# run prediction from the trained model svm1. tpred is the output of prediction
tpred = predict(svm1, tx)
```

Step 5: Plot the prediction results with the real values

```R
# Plot all the points of (prediction value, real value)
dat <- data.frame(test$Z_speed, tpred)
```
Step 6: Analyze the prediction results with the real values

```r
# correlation between the predicted values and real values
testc = cor(tpred, test$Z_speed)
# compute Relative standard error
error <- test$Z_speed - tpred
rse <- sqrt(mean(error^2))
# compute Relative percent error
errorp <- error/test$Z_speed
rsep <- sqrt(mean(errorp^2))
```

The results are compared in Figure 19. Each point represents the pair of real value and predicted value by svm1 at each hour in the 25% test data set. A regression line is drawn, this line is close to the line of \( y=x \), an ideal symmetric line. The correlation between the prediction results with the real values in the test dataset is 0.89. The RSE is 3.44 km/hour. A percent error is of 16% indicates that the prediction is valid. In summary, the average hourly travel speed of Zhedalu during the day was predicted by using the travel speed from Shangtanggaojialu, Gushanlu, and XixingBridge, as well as corresponding visibility and temperature.
This prediction was further refined to test the performance of different input attributes. First, attributes of visibility and temperature were removed from the svm prediction. This action yielded a RSE of 3.12 km/hour and percent error of 17%. Second, nighttime data (after 10:00pm and before 6:00am) was removed from the svm prediction because traffic at night is different from day time. In results, the RSE is 3.04 km/hour and the relative percent error is 16%. The accuracy of prediction is not improved by using more specific data.

SVM was applied to train and predict location based taxi pickups, a feature used to describe the pattern of people taking taxis at these locations. The hourly taxi pickup counts of the four streets, Shangtanggaojialu, Zhedalu, Gushanlu, and XixingBridge, are used, together with
two weather attributes, visibility and temperature. Taxi pickups on Zhedalu was predicted by other attributes, using training and test datasets of the given months. A RSE of 5.04 times/hour was obtained. Compared with the average pickup of 12.5 times/hour, the resulting model has a poor performance on predicting taxi pickups. The discrete and sparse values (with many zeros in different hours) of the pickups might be the reason of the poor performance.

To solve that problem, the number of pickups was grouped into four classes: Class a (pickups < 5), Class b (5\leq\text{pickups} < 10), Class c (10\leq\text{pickups} < 15), and Class d (\text{pickups}\geq 15). The svm classification was utilized to predict which class of pickups takes place on Zhedalu, using the class of pickups on other three streets, as well as corresponding weather information. Table 7 shows the matching table of the four classes on the test dataset. For values on the diagonal cells, the numbers show the correct predictive classification, while the values on the non-diagonal cells represent classification errors. The correct classification reaches 58\%, suggesting that classification improves the accuracy of prediction on hourly taxi pickups at Zhedalu.
Table 7. Matching of Predicted and Real Pickups in Four Classes

<table>
<thead>
<tr>
<th>Predicted/Real Classes</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>21</td>
<td>3</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>3</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td>d</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>57</td>
</tr>
</tbody>
</table>

**Conclusion and Future Work:** I presented a data analytics approach for pattern-based analysis of taxi trajectory data and weather data. Focusing on the correlation analysis, the results included in this thesis indicates that there is no strong correlation between taxi service and weather condition. However, this conclusion may not reflect the real situation. Two factors may contribute to it. The given three months are in winter, and the weather condition did not vary much, most of the day were either foggy or cloudy. More comprehensive data points of weather may benefit the data analysis tasks. The traffic modes of resident during business days are different from weekend. It is feasible to perform the analysis separately, data of week days and weekend. I plan to extend the data analytics approach by improved data structure, more data points of weather type, and data of more specific period.
REFERENCES


TKDE.2012.153

behavior,” in Proceedings of the 2013 ACM International Joint Conference on Pervasive

[21] D. Pfoser, S. Brakatsoulas, P. Brosch, M. Umlauft, N. Tryfona, and G. Tsironis,
“Dynamic travel time provision for road networks,” in Proceedings of the 16th ACM
SIGSPATIAL International Conference on Advances in Geographic Information Systems,
http://doi.acm.org/10.1145/1463434.1463513

[22] P. Wang, T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González,
“Understanding Road Usage Patterns in Urban Areas,” Scientific Reports, vol. 2, no.

[Online]. Available: http://dx.doi.org/10.1038/srep01376

[24] A. Thiagarajan, L. Ravindranath, K. LaCurts, S. Madden, H. Balakrishnan, S. Toledo,
and J. Eriksson, “Vtrack: Accurate, energy-aware road traffic delay estimation using
mobile phones,” in Proceedings of the 7th ACM Conference on Embedded Networked
Available: http://doi.acm.org/10.1145/1644038.1644048

38


[39] H. Wang, H. Zou, Y. Yue, and Q. Li, “Visualizing hot spot analysis result based on
mashup,” in Proceedings of the 2009 International Workshop on Location Based Social
Available: http://doi.acm.org/10.1145/1629890.1629900

[40] T. Crnovrsanin, C. Correa, C. W. Muelder, and K.-L. Ma, “Proximity-based visualization
of movement trace data,” in IEEE VAST, October 2009, pp. 11–18.

[41] Y. Gao, P. Xu, L. Lu, H. Liu, S. Liu, and H. Qu, “Visualization of taxi drivers’ income

visualization of trajectory attribute data,” IEEE Transactions on Visualization and

Conference on Visual Analytics Science and Technology (VAST), ser. VAST ’12, 2012,
pp. 183–192.

based on trajectory data,” IEEE Transactions on Visualization and Computer Graphics,


New York City Taxi Data, http://publish.illinois.edu/dbwork/open-data/.


