DO TAXI DRIVERS CHOOSE THE SHORTEST ROUTES?

A Large-Scale Analysis of Route Choice Behavior of Taxi Drivers Using Floating Car Data in Shanghai

Junyan Li

Supervisors:
Juliane Cron M.Sc. (Technische Universität München)
Univ.Prof. Mag.rer.nat. Dr.rer.nat Georg Gartner (Technische Universität Wien)
Dr. Haosheng Huang (Universität Zürich)
| 1  | Introduction               |
| 2  | Theoretical and Technical Background |
| 3  | Data and Study Area       |
| 4  | Pre-processing            |
| 5  | Overall Patterns          |
| 6  | Comparison Analysis       |
| 7  | Preference Analysis       |
| 8  | Conclusion, Limitation and Outlook |
Part 1

INTRODUCTION
1. Introduction

Context:

• **Route choice** is one of the fundamental features of **human wayfinding** (Golledge, 1995)

• Understanding route choice behavior of drivers:
  • urban planning
  • traffic management
  • infrastructure construction
1. Introduction

Context:

• **Following the shortest route** is traditionally believed to be one of the major features of human route choice behavior. (Golledge, 1995).

• Compared to **ordinary drivers**, **taxi drivers** are more experienced in wayfinding (Yao et al., 2013).
1. Introduction

Motivation:

• Several questions need to be addressed:
  • humans decision making process: rational + irrational components (de Carvalho et al., 2016).
    • makes accurate prediction of route selection difficult.
  
  • While theories predict that multiple equilibriums that result in choosing different routes (Smith, 1979, 1983, 1984),
    • hard to justify the theory prediction without large scale observation data
  
  • observation data: GPS-based point data.
    • needs pre-processing and integration with road network (Guo and Huang, 2016)

• availability of large scale Floating Car Data (FCD)
Research questions:

• To what extent do real routes overlap with the shortest routes, and how much longer are real routes compared with the shortest routes?

• Are other factors (e.g. speed limitation, road type, number of left turns, number of intersections, road usage frequency) impacting the deviation of real routes from the shortest routes?
Introduction

Importance:

• most of the existing studies focus on the behavior of ordinary drivers
• to provide cross-validation for previous studies with large scale empirical data
• To help better understanding of the patterns of route choice.
Part 2

THEORETICAL & TECHNICAL BACKGROUND
Route choice behavior

- **Rationality and bounded rationality**
  - Rationality is the fundamental assumption in many models of drivers' route choice behavior (Miller-Hooks and Mahmassani, 2000; Razo and Gao, 2013; Smith, 1979, 1983, 1984).
  - Irrationality component of drivers' evaluation also adds to the complexity of the problem. (de Carvalho et al., 2016; Guo and Huang, 2016).
- **Travel distance** (Wardrop, 1952; Tang et al., 2016; Ciscal-Terry et al., 2016; Manley et al., 2015).
- **Travel time and congestion** (Prato et al., 2014; Prato and Rasmussen, 2016).
- **Road type** (Manley et al., 2015).
Floating Car Data:

- **GPS-based FCD** is one of the reliable traffic data sources (Paulin and Bessler, 2013)
- It provides detailed trip information, such as:
  - longitude, latitude
  - time
  - speed
  - direction
- **Advantages:**
  - FCD reveals much more information, compared to survey data
  - stores information of massive number of taxi drivers with relatively cheap costs.
  - allows to systematically examine the route choice behavior of taxi drivers.
Floating Car Data:

• At the **macro** level, FCD is used to construct real time traffic information system as well as inference of mobilities and congestions. (e.g. Schafer et al., 2002; Kerner et al., 2005; Huber et al., 1999 and Gühnemann et al., 2004)

• At the **micro** level, FCD is used as a way to predict various information from individual drivers. (e.g. Simroth and Zähle, 2011; Tang et al., 2016; Yao et al., 2013; Manley et al., 2015)
Map matching: Map matching is the process of integrating the GPS-based point with spatial road network data to identify the location of a vehicle on the road network. Quddus et al. (2007) provide an overview of algorithms.
Map matching:

• **geometric-based algorithms**
  • point-to-point matching (Bentley and Maurer, 1980; Bernstein and Kornhauser, 1998)
  • point-to-curve matching (White et al., 2000)
  • curve-to-curve matching (Phuyal, 2002; White et al., 2000)

• **Limitations:**
  • based on the geometric distances between the GPS points and the road network.
  • sensitive to road network density and are unstable with presence of outliers
Map matching:

- **topological analysis**
  - take into account of the topological features of road networks
  - geometry
  - connectivity
  - contiguity
  - commonly adopt a weighted scheme that utilizing available information including *road turn, travel speed, travel direction*, and *similarity* between the vehicle trajectory and the road network (Greenfeld, 2002; Quddus et al., 2003; Yu, 2006).
Theoretical & technical background

Map matching:

• More advanced: e.g. *fuzzy logic and Bayesian analysis* (El Najjar and Bonnifait, 2005; Krakiwsky et al., 1988; Kim, 1998; Newson and Krumm, 2009)

• regards the GPS points as a serial revealing of *latent positions* on the road network that transfer between different states from previous to current time (Newson and Krumm, 2009).

• **Hidden Markov Model**: The probability of a set of GPS points located on certain combinations of road segments.
  • nearest early position
  • transferring probability matrix

• won't be impacted too much by outliers.
Theoretical & technical background

Algorithms for finding the shortest route:

• **Shortest path:**
  • The problem to find the shortest path by constructing a tree with minimum total length between n node (Dijkstra, 1959)

• **K shortest path:**
  • Given the need to compare a set of routes between two end nodes, the k-shortest path algorithm was developed to find the specified number of shortest routes from the road network in an order from 1 to k. (Hoffman and Pavley, 1959; Yen, 1971)
  • Wilson (1996) introduces the algorithm with additional weighting scheme.
State of the art and research gaps

• Yao et al. (2013) – FCD Data in Beijing
  • Influencing factors: faster travel speed, less left turns and more proportion of highways.
  • a very small subsample of taxi trips (in total 221 trips)
State of the art and research gaps

• **Yao et al. (2013) – FCD Data in Beijing**
  - Influencing factors: faster travel speed, less left turns and more proportion of highways.
  - a very small subsample of taxi trips (in total 221 trips)

• **Sun et al. (2014) – FCD Data in Shenzhen**
  - Influencing factors: travel distance, travel time and road preference
  - total number of trips is around 4000
State of the art and research gaps

• **Yao et al. (2013) – FCD Data in Beijing**
  • Influencing factors: faster travel speed, less left turns and more proportion of highways.
  • a very small subsample of taxi trips (in total 221 trips)

• **Sun et al. (2014) – FCD Data in Shenzhen**
  • Influencing factors: travel distance, travel time and road preference
  • total number of trips is around 4000

• **Manley et al. (2015) – FCD Data in London**
  • Influencing factors: travel time, turns and road category
  • total number of trips is 700,000 with about 3000 drivers,
  • limitations:
    • instead of analyzing the whole London network, the authors only select cases in smaller regions, which may not be representing the whole city.
    • the number of trips in their study is still small than real situations.
Part 3
DATA & STUDY AREA
Area of Study:
## Data & Study Area

### Data Structure:

<table>
<thead>
<tr>
<th>Column</th>
<th>Attributes</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Date</td>
<td>Date of FCD Data</td>
<td>20100615</td>
</tr>
<tr>
<td>Taxi id</td>
<td>taxi id, consists 5 letters</td>
<td>10003</td>
</tr>
<tr>
<td>Lon</td>
<td>longitude of current location</td>
<td>121.529558</td>
</tr>
<tr>
<td>Lat</td>
<td>latitude of current location</td>
<td>31.23496</td>
</tr>
<tr>
<td>Velocity</td>
<td>instant speed</td>
<td>45.5</td>
</tr>
<tr>
<td>Direction</td>
<td>driving orientation (valid number from 0 to 359)</td>
<td>155</td>
</tr>
<tr>
<td>Passenger</td>
<td>if the taxi is with passenger or not (No - 0, Yes – 1)</td>
<td>1</td>
</tr>
<tr>
<td>Time</td>
<td>exact time of current location</td>
<td>2010-06-15 09:10:05</td>
</tr>
</tbody>
</table>
Valid number of points and routes for each day included in the analysis:

Around 3,6 million valid trips, and 650 million FCD Points
Part 4

PRE-PROCESSING
Pre-processing

• Identification of OD pairs
• Filtering of irrelevant points
• Map Matching
Pre-processing

- Identification of OD pairs
- Filtering of irrelevant points
- Map Matching
Pre-processing

Identification of OD pairs

<table>
<thead>
<tr>
<th>Taxi driver ID</th>
<th>Time</th>
<th>Passenger</th>
<th>Lon</th>
<th>Lat</th>
<th>Pick-up (origin)</th>
<th>Drop-off (destination)</th>
<th>Pick-up (origin)</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1010</td>
<td>00:00</td>
<td>0</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1010</td>
<td>00:10</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1010</td>
<td>00:20</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1010</td>
<td>00:30</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1010</td>
<td>00:40</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1010</td>
<td>00:50</td>
<td>0</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1010</td>
<td>00:60</td>
<td>0</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1010</td>
<td>01:10</td>
<td>0</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1010</td>
<td>01:20</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1010</td>
<td>01:30</td>
<td>1</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Pre-processing

• Identification of OD pairs

• Filtering of irrelevant points

• Map Matching
Pre-processing

Filtering of irrelevant points:

• Why?
Pre-processing

Filtering of irrelevant points:

- Why?

  - taxi driver may forget to turn off the passenger-on-board indicator when the taxi is off-service and not moving.
Pre-processing

Filtering of irrelevant points:

- **Why?**
  - taxi driver may forget to turn off the passenger-on-board indicator when the taxi is off-service and not moving.
  - taxi drivers may mistakenly turn off or turn on the passenger-on-board indicator in a very short time.
Pre-processing

- Identification of OD pairs
- Filtering of irrelevant points
- Map Matching
Pre-processing

Map Matching:

Hidden Markov Model (HMM) (Newson and Krumm, 2009)

Consider:

- recorded GPS locations
- timestamp of recorded points
- connectivity of the road network

Calculating the path with the highest likelihood
Pre-processing

Map Matching:
Pre-processing

Map Matching:
Part 5

OVERALL PATTERNS
Overall Patterns

• Distribution of service areas

• Taxi service activities

• Real routes
Overall Patterns

Distribution of service areas:

centered around city center
Overall Patterns

Distribution of service areas:

centered around city center

Density of the drop-off locations
Overall Patterns

- Distribution of service areas
- Taxi service activities
- Real routes
Overall Patterns

Taxi service activities:

- Regularity of taxi services

Event plot of pick-up activities across days
Overall Patterns

Taxi service activities:

• Regularity of taxi services
Overall Patterns

Taxi service activities:

• Service duration

  Distribution of travel duration

  Mean: 14.54 min
Overall Patterns

Taxi service activities:

- Service distance

  Distribution of travel distance of on-service taxi drivers

  Mean: 7.27km
Overall Patterns

• Distribution of service areas
• Taxi service activities
• Real routes
Overall Patterns

Real routes:

a) Map-matched real routes
Overall Patterns

Real routes:

b) Density of the routes
Part 6

COMPARISON ANALYSIS
Comparison Analysis

- Route-based Analysis
  - Percentage of shared length (PSL)
  - Percentage of length difference (PLD)
  - Route Circuity

- Route segment-based analysis
  - Road type
  - Intersections and number of turns
Comparison Analysis

- **Route-based Analysis**
  - Percentage of shared length (PSL)
  - Percentage of length difference (PLD)
  - Route Circuity

- **Route segment-based analysis**
  - Road type
  - Intersections and number of turns
Percentage of shared length (PSL):

- To investigate the extent of similarity between the real and the shortest routes.
- The PSL is calculated by first identifying the overlapped segments of real routes and the shortest routes and then dividing the overlapped length by real total length.

\[
PSL = \frac{\sum_{i \in \Phi} r_i}{\sum_{j \in \Theta} r_j}
\]

where \( \Phi \) and \( \Theta \) denote the set of overlapped segments and the segments contained in real routes, \( i \) denotes individual segments, and \( r_i \) denotes the segment length.
Comparison Analysis

Percentage of shared length (PSL) :

• Overall result

<table>
<thead>
<tr>
<th>PSL</th>
<th>Percentage (%)</th>
<th>Cumulative percentage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.66727</td>
<td>16.66727</td>
</tr>
<tr>
<td>0.8 -1</td>
<td>5.54906</td>
<td>22.21633</td>
</tr>
<tr>
<td>0.5 - 0.8</td>
<td>19.64954</td>
<td>41.86587</td>
</tr>
<tr>
<td>0.2 - 0.5</td>
<td>26.71682</td>
<td>68.58269</td>
</tr>
<tr>
<td>0 - 0.2</td>
<td>31.41731</td>
<td>100</td>
</tr>
</tbody>
</table>

Average: 0.4565

Table 4. Distribution of PSL
Comparison Analysis

Percentage of shared length (PSL):

• Variance between different time periods

Figure 14. PSL grouped by time slots
Comparison Analysis

Percentage of shared length (PSL):

• Variance between different time periods

Figure 15. PSL grouped by weekdays and weekends
Comparison Analysis

Percentage of shared length (PSL):

- Variance between different travel distances

Figure 16. PSL grouped by lengths of real routes
Comparison Analysis

- Route-based Analysis
  - Percentage of shared length (PSL)
  - Percentage of length difference (PLD)
  - Route Circuity

- Route segment-based analysis
  - Road type
  - Intersections and number of turns
Comparison Analysis

Percentage of length difference (PLD)
• PLD compares the total length between the real and the shortest routes to show how much longer is the actual route to the shortest routes.

\[
PLD = \frac{\sum_{i \in \Gamma} r_i - \sum_{j \in \Delta} r_j}{\sum_{i \in \Gamma} r_i}
\]

where \( \Gamma \) and \( \Delta \) denote respectively the set of real route segments and the shortest route segments, \( i \) denotes an individual segment, and \( r_i \) and \( r_j \) denotes the length of the segment. Or in short: \( PLD = \frac{L_1 - L_2}{L_1} \), which means the value is calculated by subtracting the length of the shortest route (\( L_2 \)) from the real route (\( L_1 \)) and then divided by the length of the real route.
Comparison Analysis

Percentage of length difference (PLD)

• Overall result

<table>
<thead>
<tr>
<th>PLD</th>
<th>Percentage (%)</th>
<th>Cumulative percentage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.8 -1</td>
<td>0.6421462</td>
<td>0.6421462</td>
</tr>
<tr>
<td>0.5 - 0.8</td>
<td>4.2816988</td>
<td>4.923845</td>
</tr>
<tr>
<td>0.2 - 0.5</td>
<td>26.540425</td>
<td>31.46427</td>
</tr>
<tr>
<td>0 -0.2</td>
<td>68.53573</td>
<td>100</td>
</tr>
</tbody>
</table>

Average: 0.1609

Table 5. Distribution of PLD
Comparison Analysis

Percentage of length difference (PLD)

- Variance between different time periods

Figure 17. PLD grouped by time slots
Comparison Analysis

Percentage of length difference (PLD)
- Variance between different time periods

Figure 18. PLD grouped by weekdays and weekends
Comparison Analysis

Percentage of length difference (PLD)

• Variance between different travel lengths

Figure 19. PLD grouped by lengths of real routes
Comparison Analysis

- Route-based Analysis
  - Percentage of shared length (PSL)
  - Percentage of length difference (PLD)
  - Route Circuity

- Route segment-based analysis
  - Road type
  - Intersections and number of turns
For each OD pair $a, b$, the route circuity is calculated as follows,

$$r_{CR} = \frac{l_{ab} - d_{ab}}{l_{ab}}$$

where $l_{ab}$ refers to the route distance, and $d_{ab}$ refers to the geometric distance between point $a$ and point $b$, and

$$d_{ab} = \arccos(\sin \theta_a * \sin \theta_b + \cos \theta_a * \cos \theta_b * \cos \Delta \lambda) * R$$

where $\theta$ refers to latitude, $\lambda$ refers to longitude, $R$ refers to the radius of the earth.
Comparison Analysis

Route Circuity

• Overall result

<table>
<thead>
<tr>
<th>Route Circuity</th>
<th>Percentage in real routes (%)</th>
<th>Cumulative percentage(%)</th>
<th>Percentage in shortest routes (%)</th>
<th>Cumulative percentage(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0.8 - 1</td>
<td>1.1759</td>
<td>1.1759</td>
<td>0.09560</td>
<td>0.0956</td>
</tr>
<tr>
<td>0.5 - 0.8</td>
<td>12.1467</td>
<td>13.3225</td>
<td>1.3021</td>
<td>1.3977</td>
</tr>
<tr>
<td>0.2 - 0.5</td>
<td>62.7011</td>
<td>76.0236</td>
<td>37.6061</td>
<td>39.0037</td>
</tr>
<tr>
<td>0 - 0.2</td>
<td>23.9764</td>
<td>100</td>
<td>60.9963</td>
<td>100</td>
</tr>
</tbody>
</table>

Average
Real routes: 0.3154
Shortest routes: 0.1862

Table 6. Distribution of Route Circuity of real routes and shortest routes
Comparison Analysis

Route Circuity

• Variance between different time periods

**Figure 20.** Route circuity grouped by time slots
Comparison Analysis

Route Circuity

• Variance between different time periods

Figure 21. Route circuity grouped by weekdays and weekends
Comparison Analysis

Route Circuity

• Variance between different travel distance

Figure 22. Route circuity grouped by length of real routes
Comparison Analysis

- **Route-based Analysis**
  - Percentage of shared length (PSL)
  - Percentage of length difference (PLD)
  - Route Circuity

- **Route segment-based analysis**
  - Road type
  - Intersections and number of turns
Road type

As traditionally categorized, five types of road based on Open Street Map data are analyzed: motorway, primary, secondary, tertiary and residential roads.

• Motorway: 120 km/h
• Primary road: 70 km/h
• Secondary road: 60 km/h
• Tertiary road: 40 km/h
• Residential road: 40 km/h
**Comparison Analysis**

**Road type**

![Bar chart showing the percentage of different types of roads in real and shortest routes.]

**Figure 23.** Share of different types of roads in the real and shortest routes.
Comparison Analysis

Road type

Figure 24. Share of different types of roads across travel distances
Comparison Analysis

Road type

Figure 25. Share of different types of roads across time slots
Comparison Analysis

Road type

Figure 26. Share of different types of roads in weekdays and weekends
Comparison Analysis

- **Route-based Analysis**
  - Percentage of shared length (PSL)
  - Percentage of length difference (PLD)
  - Route Circuity

- **Route segment-based analysis**
  - Road type
  - Intersections and number of turns
Comparison Analysis

Intersections and number of turns

- Right-hand traffic in China
- left turns (210° – 330°)
- right turns (30° - 150°)
- The straight ways are the ways whose nodes are neither the left- nor the right-turns, or the ways that only connect with two nodes.
Comparison Analysis

Intersections and number of turns

Figure 27. Comparison of numbers of straight ways in the real and the shortest routes
Comparison Analysis

Intersections and number of turns

Figure 28. Comparison of numbers of left turns in the real and the shortest routes
Comparison Analysis

Intersections and number of turns

Figure 30. Comparison of numbers of intersections in the real and the shortest routes
Part 7

PREFERENCE ANALYSIS
- Scenarios and their application to general routes
- Scenarios applied to airport-city center routes
Preference Analysis

Scenarios and their application to general routes

Scenario 0: Minimal distance
Preference Analysis

Scenarios and their application to general routes

Scenario 0: Minimal distance
Scenario 1: Least free flow travel time
Preference Analysis

Scenarios and their application to general routes

- Scenario 0: Minimal distance
- Scenario 1: Least free flow travel time
- Scenario 2: Least observed travel time
Preference Analysis

Scenarios and their application to general routes

Scenario 0: Minimal distance
Scenario 1: Least free flow travel time
Scenario 2: Least observed travel time
Scenario 3: Minimal (road category * distance)
Scenario 4: Minimal (road category * time)
Scenarios and their application to general routes

Scenario 0: Minimal distance
Scenario 1: Least free flow travel time
Scenario 2: Least observed travel time
Scenario 3: Minimal (road category * distance)
Scenario 4: Minimal (road category * time)
Scenario 5: Minimal (road usage frequency * distance)
Scenario 6: Minimal (road usage frequency * time)
Preference Analysis

Scenarios and their application to general routes

Scenario 0 : Minimal distance
Scenario 1 : Least free flow travel time
Scenario 2 : Least observed travel time
Scenario 3 : Minimal (road category * distance)
Scenario 4 : Minimal (road category * time)
Scenario 5 : Minimal (road usage frequency * distance)
Scenario 6 : Minimal (road usage frequency * time)
Scenario 7 : Minimal (left turns * distance)
Scenario 8 : Minimal (left turns * time)
Preference Analysis

Scenarios and their application to general routes

Scenario 0 : Minimal distance
Scenario 1 : Least free flow travel time
Scenario 2 : Least observed travel time
Scenario 3 : Minimal (road category * distance)
Scenario 4 : Minimal (road category * time)
Scenario 5 : Minimal (road usage frequency * distance)
Scenario 6 : Minimal (road usage frequency * time)
Scenario 7 : Minimal (left turns * distance)
Scenario 8 : Minimal (left turns * time)
Scenario 9 : Minimal (turns * distance)
Scenario 10 : Minimal (turns * time)
## Preference Analysis

### Scenarios and their application to general routes

<table>
<thead>
<tr>
<th>Factors</th>
<th>S0</th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
<th>S7</th>
<th>S8</th>
<th>S9</th>
<th>S10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Free flow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed travel time</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road type</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road use frequency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Left turns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Turns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Table 7.* Scenarios: the gray cells indicate the inclusion of the factor into the corresponding scenario.
Scenarios and their application to general routes

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Weighted Criteria</th>
<th>mean(%)</th>
<th>Over 75% match</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Minimal distance</td>
<td>46.61</td>
<td>25.69</td>
</tr>
<tr>
<td>1</td>
<td>Least freeflow travel time</td>
<td>51.30</td>
<td>30.77</td>
</tr>
<tr>
<td>2</td>
<td>Least observed travel time</td>
<td>51.75</td>
<td>30.92</td>
</tr>
<tr>
<td>3</td>
<td>Minimal (road category * distance)</td>
<td>48.79</td>
<td>29.39</td>
</tr>
<tr>
<td>4</td>
<td>Minimal (road category * time)</td>
<td>45.80</td>
<td>26.72</td>
</tr>
<tr>
<td>5</td>
<td>Minimal (road usage frequency * distance)</td>
<td>53.56</td>
<td>32.52</td>
</tr>
<tr>
<td>6</td>
<td>Minimal (road usage frequency * time)</td>
<td>52.48</td>
<td>31.94</td>
</tr>
<tr>
<td>7</td>
<td>Minimal (left turns * distance)</td>
<td>46.61</td>
<td>25.60</td>
</tr>
<tr>
<td>8</td>
<td>Minimal (left turns * time)</td>
<td>51.06</td>
<td>30.50</td>
</tr>
<tr>
<td>9</td>
<td>Minimal (turns * distance)</td>
<td>36.23</td>
<td>14.52</td>
</tr>
<tr>
<td>10</td>
<td>Minimal (turns * time)</td>
<td>38.46</td>
<td>16.53</td>
</tr>
</tbody>
</table>

Table 8. Prediction accuracy of different scenarios (cells show the cumulative percentage of cases within specified accuracy category)
Preference Analysis

- Scenarios and their application to general routes
- Scenarios applied to airport-city center routes
Scenarios applied to airport-city center routes

• **DBSCAN**
  • *Density-Based Spatial Clustering of Applications with Noise*
  • to identify the clusters of points (Ester et al., 1996)
  • two parameters:
    • the minimum number of points (minpts)
    • the range of clustering ($\varepsilon$)
Preference Analysis

Scenarios applied to airport-city center routes

- DBSCAN

<table>
<thead>
<tr>
<th>N</th>
<th>100</th>
<th>200</th>
<th>300</th>
<th>400</th>
<th>500</th>
<th>600</th>
<th>700</th>
<th>800</th>
<th>900</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>4227</td>
<td>1522</td>
<td>558</td>
<td>147</td>
<td>53</td>
<td>36</td>
<td>27</td>
<td>18</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>200%</td>
<td>3204</td>
<td>1201</td>
<td>400</td>
<td>99</td>
<td>44</td>
<td>36</td>
<td>27</td>
<td>19</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>300%</td>
<td>2885</td>
<td>929</td>
<td>296</td>
<td>77</td>
<td>36</td>
<td>29</td>
<td>18</td>
<td>15</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>400%</td>
<td>2107</td>
<td>735</td>
<td>256</td>
<td>59</td>
<td>27</td>
<td>18</td>
<td>12</td>
<td>10</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>500%</td>
<td>1746</td>
<td>594</td>
<td>206</td>
<td>70</td>
<td>24</td>
<td>11</td>
<td>10</td>
<td>8</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>600%</td>
<td>1455</td>
<td>487</td>
<td>180</td>
<td>68</td>
<td>24</td>
<td>15</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>700%</td>
<td>1227</td>
<td>409</td>
<td>154</td>
<td>66</td>
<td>28</td>
<td>15</td>
<td>7</td>
<td>9</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>m</td>
<td>1050</td>
<td>345</td>
<td>145</td>
<td>66</td>
<td>22</td>
<td>18</td>
<td>11</td>
<td>7</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>n</td>
<td>900</td>
<td>302</td>
<td>125</td>
<td>61</td>
<td>28</td>
<td>13</td>
<td>7</td>
<td>5</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>p</td>
<td>774</td>
<td>267</td>
<td>103</td>
<td>59</td>
<td>26</td>
<td>15</td>
<td>10</td>
<td>6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>t</td>
<td>693</td>
<td>228</td>
<td>93</td>
<td>51</td>
<td>29</td>
<td>16</td>
<td>10</td>
<td>8</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>s</td>
<td>606</td>
<td>203</td>
<td>86</td>
<td>51</td>
<td>24</td>
<td>15</td>
<td>9</td>
<td>8</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>518</td>
<td>185</td>
<td>87</td>
<td>43</td>
<td>22</td>
<td>11</td>
<td>11</td>
<td>7</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>1.5</td>
<td>453</td>
<td>158</td>
<td>77</td>
<td>41</td>
<td>20</td>
<td>13</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>391</td>
<td>145</td>
<td>81</td>
<td>34</td>
<td>29</td>
<td>12</td>
<td>11</td>
<td>10</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>2.5</td>
<td>352</td>
<td>130</td>
<td>79</td>
<td>35</td>
<td>25</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>322</td>
<td>124</td>
<td>76</td>
<td>33</td>
<td>21</td>
<td>11</td>
<td>11</td>
<td>9</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>3.5</td>
<td>282</td>
<td>104</td>
<td>68</td>
<td>29</td>
<td>19</td>
<td>15</td>
<td>8</td>
<td>8</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>259</td>
<td>88</td>
<td>61</td>
<td>26</td>
<td>18</td>
<td>15</td>
<td>7</td>
<td>7</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>4.5</td>
<td>229</td>
<td>83</td>
<td>58</td>
<td>30</td>
<td>16</td>
<td>14</td>
<td>8</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

$\epsilon = 500$

$\text{minpts} = 3000$

Table 9. Number of cluster with different combinations of parameters (minpts in the first column and \( \epsilon \) in the first row)
Preference Analysis

Scenarios applied to airport-city center routes

• Results

Figure 31. Clusters of most frequently visited places
Preference Analysis

Scenarios applied to airport-city center routes

• Results

Figure 32. Clusters of most frequently visited places by routes from the Pudong International Airport.
Preference Analysis

Scenarios applied to airport-city center routes

• Results

Figure 33. Routes from the Pudong International Airport to city center
Preference Analysis

Scenarios applied to airport-city center routes

• Results

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Weighted Criteria</th>
<th>Mean (%)</th>
<th>Over 75% match</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Minimal distance</td>
<td>13.25</td>
<td>0.00</td>
</tr>
<tr>
<td>1</td>
<td>Least freeflow travel time</td>
<td>24.22</td>
<td>4.67</td>
</tr>
<tr>
<td>2</td>
<td>Least observed travel time</td>
<td>48.00</td>
<td>12.62</td>
</tr>
<tr>
<td>3</td>
<td>Minimal (road category * distance)</td>
<td>51.79</td>
<td>17.76</td>
</tr>
<tr>
<td>4</td>
<td>Minimal (road category * time)</td>
<td>30.78</td>
<td>9.35</td>
</tr>
<tr>
<td>5</td>
<td>Minimal (road usage frequency * distance)</td>
<td>32.34</td>
<td>0.00</td>
</tr>
<tr>
<td>6</td>
<td>Minimal (road usage frequency * time)</td>
<td>25.84</td>
<td>5.14</td>
</tr>
<tr>
<td>7</td>
<td>Minimal (turns * distance)</td>
<td>10.68</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>Minimal (turns * time)</td>
<td>20.52</td>
<td>4.67</td>
</tr>
<tr>
<td>9</td>
<td>Minimal (left turns * distance)</td>
<td>13.26</td>
<td>0.00</td>
</tr>
<tr>
<td>10</td>
<td>Minimal (left turns * time)</td>
<td>24.20</td>
<td>4.66</td>
</tr>
</tbody>
</table>

Table 10. Prediction accuracy of different scenarios to the routes from Pudong International Airport to the city center
Part 8
CONCLUSION, LIMITATION & OUTLOOK
Conclusion:

Real routes to some extent overlap with the shortest routes with an average value of 46%.
Conclusion:

Real routes to some extent overlap with the shortest routes with an average value of 46%.

Other factors, especially the road type, road usage frequency, speed limit and observed travel speed, have impacts on the route choices.
Conclusion, Limitation & Outlook

Conclusion:

In the general situation, the best scenario in combination with travel distance and road usage frequency predicts 53.56% of real routes compare to 46.61% in the shortest routes.
Conclusion: 

In the **general** situation, the best scenario in combination with **travel distance** and **road usage frequency** predicts **53.56%** of real routes compare to **46.61%** in the shortest routes.

In **long distant** trips, the scenario with **road type** and **travel distance** predicts **51.79%** of real routes compared to **13.25%** in the shortest routes.
Limitation & Outlook:

• due to computational constraints, only the top 10 routes to predict the best matches in Scenario 7, 8, 9 and 10 could be calculated for each real routes.
Limitation & Outlook:

• due to computational constraints, only the top 10 routes to predict the best matches in Scenario 7, 8, 9 and 10 could be calculated for each real routes.
  • to optimize the efficiency of the k-shortest-path algorithm so that it is computable with large-scale road networks.
Conclusion, Limitation & Outlook

Limitation & Outlook:

• due to computational constraints, only the top 10 routes to predict the best matches in Scenario 7, 8, 9 and 10 could be calculated for each real routes.
  • to optimize the efficiency of the k-shortest-path algorithm so that it is computable with large-scale road networks.

• period of study
  • extend the period of study and examine the impact of other external events on the taxi drivers' behavior.
Limitation & Outlook:

• due to computational constraints, only the top 10 routes to predict the best matches in Scenario 7, 8, 9 and 10 could be calculated for each real routes.
  • to optimize the efficiency of the k-shortest-path algorithm so that it is computable with large-scale road networks.

• period of study
  • extend the period of study and examine the impact of other external events on the taxi drivers' behavior.

• context aware analysis
  • extend the analysis to investigate the impact on context


THANKS!