Modeling of Cyclists Acceleration Behavior Using Naturalistic Data

DING LUO
Modeling of Cyclists Acceleration Behavior Using Naturalistic Data

DING LUO

Master of Science Thesis
Department of Transport Science
Division of Traffic and Logistics
Stockholm, Sweden 2014

TSC-MT 14-007
Abstract

Over the past few years, many cities have witnessed the increasing popularity of cycling, especially among ordinary commuters. Accordingly, there has also been a fast growing demand for the knowledge of cycling performance as well as cyclist behavior, which can be valuable for both traffic planners and policy makers when it comes to the bicycle-related issues. The aim of this study, hence, is to investigate the cycling performance in detail and to further develop proper models which can be implemented in the microscopic cycling traffic simulation.

The study was initiated with data collection in the summer of 2013 in Stockholm. A number of commuter cyclists were recruited and then provided with GPS devices to record their daily cycling trips. The GPS devices were portable but qualified enough to measure cyclists' position, speed and altitude with a time interval of one second. Before the winter, around 100 natural cycling trips made in the urban area of Stockholm were collected and a database was later established to manage the raw data. Prior to the data analysis, measurement noise cancellation and profile smoothing were performed by implementing multiple processing approaches, including the robust locally weight regression and the Kalman filtering. A cycling regime which separates the cyclist behavior into three different kinds (acceleration, deceleration and cruising) was constructed based on the data observation. According to this regime, a normal cyclist should always endeavor to achieve and maintain a desired speed which varies depending on a number of factors, such as the cyclist’s own demographics and the road grade. If a cyclist’s present speed was not corresponding to her present desired speed, she would accelerate or decelerate immediately. Based on this assumption, the GPS data were classified into three parts, including dedicated datasets for acceleration profiles, deceleration profiles and cruising profiles. The profiles were analyzed statistically and some significant cycling characteristics were founded. Moreover, mathematical models were formulated to describe cyclists’ acceleration and deceleration behavior. The models were further estimated using the maximum likelihood estimator and evaluated by several goodness-of-fit measures.
Acknowledgements

It’s been a journey, not long, but fantastic, and eventually, I am going to finish it with this trivial achievement. My first deep gratitude goes to Xiaoliang Ma, my supervisor, who guided me to this fabulous area of transport research and has kept helping and supporting me to make progress. Without the first opportunity offered by him of undertaking an iTRUE project, I could not have improved myself so much over the past ten months. I would also like to express my gratitude to the people in the division of Traffic and Logistics, where I did my thesis. I will never forget those Tuesdays when we together played Innebandy, an indoor floorball game which is very popular in Sweden. Special thanks go to three PhD students in the division: Mahmood, Athina and Zifei. Thank you for sharing rooms with me and helping me all the time.

I want to express my deep gratitude to all my friends, classmates, and especially, the people in the English debate club at Cafe String. Thanks for having accompanied me over the last two years. Life in Sweden would have been too boring if you had not been around.

Finally, and perhaps most importantly, I want to express my deepest gratitude to my family. Without the constant support from you, I would have no chance to even start this adventure in Sweden.

Ding Luo
Stockholm, June 2014
Contents

1 Introduction ................................................. 1
   1.1 Background ........................................... 1
   1.2 Motivation and Objectives .............................. 2
   1.3 Literature Review ...................................... 3
   1.4 Thesis Outline and Contributions ...................... 6

2 Data Preparation ........................................ 7
   2.1 Data Collection ........................................ 7
   2.2 Preprocessing of Altitude Data .......................... 8
      2.2.1 Problem Statement ................................ 8
      2.2.2 Methodology ...................................... 10
      2.2.3 Practical Results .................................. 11
   2.3 Preprocessing of GPS Data .............................. 14
      2.3.1 Problem Statement ................................ 14
      2.3.2 Methodology ...................................... 14
      2.3.3 Practical Results .................................. 18
   2.4 Acceleration Profile Estimation .......................... 20
   2.5 Gradient Profile Estimation ............................. 21
   2.6 Summary .............................................. 21

3 Model Development ....................................... 23
   3.1 The Cycling Regime .................................... 23
   3.2 The Cycling Performance ................................ 24
      3.2.1 Profile Selection ................................... 24
      3.2.2 Profile Analysis .................................... 27
### CONTENTS

3.3 The Acceleration Model .................................. 34
  3.3.1 The Polynomial Model ................................ 34
  3.3.2 The Acceleration Model .................................. 35
  3.3.3 Likelihood Function Formulation ......................... 36
  3.3.4 Model Specification ...................................... 36
  3.3.5 Estimation Results ....................................... 38
  3.3.6 Model Validation .......................................... 39
3.4 The Deceleration Case .................................. 40
3.5 Discussion ............................................. 42
3.6 Summary ............................................... 43

4 Conclusions and Future Work .......................... 45
  4.1 Conclusions ............................................. 45
  4.2 Future Work ............................................ 46

A Acceleration and Deceleration Profile Illustration ........ 49

B MATLAB Code for Model Estimation and Validation .......... 57

Bibliography .................................................. 63
Chapter 1

Introduction

1.1 Background

Invented in the 19th century, the first modern chain-driven bicycle marks the beginning of a great era in which human’s mobility is greatly enhanced. As a simple but efficient transportation tool, the bicycle used to be important for people’s daily life, making commuter trips much more effortless, yet more enjoyable than other transportation modes. However, due to the later great development of the automotive industry, private cars soon became accessible to thousands of millions of individuals and families, therefore causing a constant decrease in the number of commuter cyclists over decades. People then started enjoying the process of shifting from being a cyclist to being a driver instead. As a result, bicycles were forgotten by most people and gradually become leisure tools for entertainment.

Massive issues arise, nevertheless, with the tremendous increase in both the car ownership and car usage over the second half of the 20th century. Among all problems, traffic congestion turns our to be one of the most severe ones. Commuters around the world started suffering from the worsening traffic conditions. They spend countless hours on their commuter trips and always end up them with a terrible mood. In addition to the congestion, a large proportion of drivers have also been bothered by the fast growing expenditure of driving, such as the parking fee and fuel cost. Besides the effects on driver commuters, the overuse of vehicles has also threatened the environment and further jeopardized our health. Issues in relate to private cars, such as the green house emission, have become urgent to be addressed by the whole society.

Given the current situation, many have been seeking effective solutions to vehicle-related
problems, or available alternatives to this motorized transportation. During this process, cycling shows up as a sustainable and healthy way to travel around in urban areas. People in many counties, especially in developed ones, start riding bicycles again. For instance, Copenhagen stands out as one of the leading cities in terms of the bicycle ridership. It boasts a bicycle mode share of 35 percent and Copenhageners are now very willing to travel by bike [1]. Portland, one of the metropolitan cities in the United States, has also seen a 10-year consecutive increase in the bicycle mode share according to the US Census [2]. This phenomenon is common and a large number of people have realized that cycling are much better than driving their private cars in terms of multiple aspects. First, it allows the commuter to worry much less about the traffic condition before the departure. They would be troubled much less by traffic congestion and parking problems without driving a car. More importantly, both the environment and commuter herself benefit from the choice of cycling. On the one hand, there would be less consumption of fossil fuel, hence reducing the greenhouse gas emissions. On the other hand, the commuter can gain long-term health profits. This has be verified by a recent study [3], which also indicates that commuter cyclists are even able to live longer. With all these significant advantages of cycling, it can be firmly inferred that the upward trend of commuter cyclists, in the long term, can be irreversible in more places.

1.2 Motivation and Objectives

As described in the background, there has been a growing number of commuters re-attracted by cycling over the past few years. However, the sharp increase in the cycling community in many cities, such as Stockholm where more than 150,000 cyclists, mainly commuters, pass the inner city border every day during the warmer half of the year (mid-April to mid-October) [4], has not been matched by proportionate investments in bicycle facilities. Furthermore, there is also a lack of powerful tools which can assist in bicycle traffic planning or accident prevention, such as specialized bicycle traffic simulators.

Traffic simulation is an analytical and effective approach for researchers to investigate traffic phenomena and address traffic problems. It is known that the development of traffic simulation tools largely depends on the accumulation and advance of fundamental theories. For instance, theories which can specifically illustrate driver behavior, such as the car-following and lane-changing models, are crucial in microscopic traffic simulation because it
1.3. LITERATURE REVIEW

is exactly these theories and models that determine how vehicles in the simulator should accelerate, decelerate and interact with other objects in accordance with practical rules. With a pursuit of achieving premium car simulation, great efforts on studies into driver behavior and vehicle performance have been done by many researchers, whereas these efforts have not had a counterpart for research into cyclist behavior. Given the shortage, the author was therefore motivated to conduct a deep and systematic study into the cyclist behavior and to develop mathematical models which can be particularly valuable to the microscopic bicycle traffic simulation. The methodological framework of the study is shown in Figure. Notably, naturalistic cycling data collected by GPS devices were used in the study, which enhances both the universality and practicability of the study results. With a comprehensive study ultimately presented, the author hopes that it can contribute to the further development of cycling facilities and cyclist communities.

1.3 Literature Review

Studies on bicycle and cyclists cover a wide range of topics, including the bicycle kinematics, bicycle traffic planning, cycling safety issues and so on. However, with the emphasis of this thesis placed upon the investigation into cyclists’ natural performance, such as their acceleration and deceleration behavior, only those literatures of which topics are related to our research interest are reviewed in this section.

One of the classical ways of modeling cycling performance is to develop mathematical models depicting how the cycling speed varies depending on the pedaling power. This research trend started in the 1980s, with Martin et al.\[5\] in 1998 reaching the peak by establish a state-of-the-art model. Essentially, the so-called state-of-the-art model successfully reveals the equilibrium between the pedaling power and other resistance powers. The model is illustrated by Equation (1.1).

\[
P_{\text{ped}} + P_{\text{air}} + P_{\text{bear}} + P_{\text{roll}} + P_{\text{kin}} = \tau P_{\text{ped}} \tag{1.1}
\]
where \( P_{\text{pot}} \), \( P_{\text{air}} \), \( P_{\text{bear}} \), \( P_{\text{roll}} \) and \( P_{\text{kin}} \), respectively, denote the power from the potential energy, the aerodynamic drag, the frictional losses in wheel bearings, the rolling resistance and the kinetic energy. \( P_{\text{ped}} \), on the other side, denotes the pedaling power from the cyclist who intends to propel the bicycle and \( \tau \) denotes an efficiency factor which accounts for the frictional loss in the drive chain.

On the basis of this state-of-the-art model, a number of researchers have tried to extend and implement it to realistic studies. For instance, one of the recent studies by Dahmen et al. [6] successfully apply the model to simulation. They integrate the calibrated state-of-the-art model into an empirical cycling study in which data acquisition, analysis, modeling and simulation were all incorporated. Their results show that using the simulator which applies the state-of-the-art model leads accurate predictions of the cycling performance. Examples like this, therefore, prove the feasibility of this particular research trend.

Besides contributions of detailed mathematical models based on physical laws, there is also some cycling-related research conducted from engineering and statistical perspectives. Field measurements are always a significant part in these studies. There are mainly two types of measurement techniques used by different researchers, one of which is videotaping. As a typical method to collect traffic data, videotaping has been adopted by many researchers for a long time (see [7], [8], [9], [10]). This method requires at least one camera, but usually more to record the field condition continuously over a period. Subsequently it also demands great manual effort to review the videotapes and create the dataset. It can be found that videotaping is more often used by studies targeting events at intersections because cameras have to be set at certain spots with good angles. More recently, some studies, such as [11] and [12], have developed new processing software to extract each cyclists trajectory through an intersection, which may make videotaping more efficient and accurate.

Collecting cyclists’ GPS data is the other way to help to investigate cycling performance and cyclist behavior. This method is relatively new in the domain of cycling researches, yet has been increasingly operable thanks to the rapid development of commercial GPS techniques. The civilian GPS signal used to be intentionally degraded by the U.S. government, making it almost impossible to accurately locate a moving object by using GPS. However, this degradation was canceled in 2000 and the GPS industry began to boom from then on. Currently even the portable GPS devices can report pretty accurate position and the usage of large amount of GPS data in some recent studies, such as [13], [14], and [15], have gained success in revealing more overall and representative characteristics of ordinary cyclists than
previous studies did. For instance, Parkin et al. [14] in the U.K. found both the speed and acceleration are significantly linearly correlated with the road grade, with the mean speed and mean acceleration rate decreasing as the road grade rises from downhill to uphill. This conclusion came from substantial GPS data (547 cycling trips) collected by 16 volunteers (four female). In another study [13], El-Geneidy et al. developed regression models using real-time GPS data collected from a small sample of cyclists traveling on various types of facilities in Minneapolis, MN. They tried to account for the cycling speed by incorporating multiple variables, including trip characteristics, gender, the presence of an off-street facility, and an individual’s comfort level with traveling in heavy traffic.

Most studies have been dedicated to investigating cycling performance at intersections. This is because travelers riding a bicycle always cross unprotected zones and interact with vehicles at relatively high speeds, which increases the possibility of accidents at these spots. A previous study [16] testified this speculation, indicating that vehicle-bicycle collisions turn out to be the most common signal timing issue due to insufficient clearance time for a cyclist maintaining his cruising speed. In 1997, Pein examined cyclist performance on multiuse trails as well as at three-leg intersections in his study [8]. A linear function of distance was suggested by him, which can be used to estimate bicycle crossing velocity and acceleration based on the regression coefficients. Landis et al. in their study [11] illustrated how field-collected observations from a basic video setup can be used to estimate acceleration and speed with the application of equations of motion. Figliozzi et al. conducted a series of studies investigating cyclists’ performance at intersections in Portland, Oregon (see [10] and [17]). They found statistically significant differences between male and female in the crossing times at intersections in [17] and further developed a practical method to estimate cyclists’ acceleration and speed at intersections in [10]. It has to be pointed out all these studies assumed that cyclists accelerate or decelerate uniformly, although Pein in his study [8] mentioned that bicyclists may exhibit nonuniform acceleration. However, he simply attributes the nonuniformity to a couple of external factors, such fitness and trip purpose.

Besides the studies mentioned above, using emerging technologies to investigate cycling performance and cyclist behavior has also been a trend. For example, Dozza et al. in their latest study [15] show some novel results about bicycle dynamics and cyclist behavior based on a great amount of naturalistic field data collected in Gothenburg, Sweden. The cycling data were measured by multiple instruments, including GPS devices, brake sensors, Inertial Measurement Unit and so on. They suggest that lateral acceleration is a key piece
of information and present three ways to extract it using independent sensors embedded in inertial measurement units which allow them to later obtain a more robust estimation. Overall, this paper provides a new perspective of examining cycling behavior.

1.4 Thesis Outline and Contributions

In this section, both the contents of the thesis and its contributions are outlined.

Chapter 2: Data Preparation

This chapter first introduces the data collection performed in Stockholm during August and October in 2013. It then presents how raw GPS data were processed with the application of multiple approaches.

Chapter 3: Model Development

This chapter demonstrates all the work concerning the model development. It first introduces a cycling regime which serves as the basis of the research and then illustrates some results from the profile analysis. Afterwards, it continues to present the most important contribution of the thesis, which is the development of both acceleration and deceleration models for the cyclist behavior.

Chapter 4: Conclusion and Future Work

This chapter summarizes the findings of the current study and identifies the directions for future research.

Contributions

This thesis enlightens a prospective way of investigating cycling performance and cyclist behavior using large amount of GPS data. It provides a method to model cyclists’ acceleration and deceleration behavior based on selected acceleration and deceleration profiles. In addition, the applications of data processing approaches, such as the locally weighted regression and the Kalman filtering, can be also regarded as meaningful references for other studies which intend to deal with discrete time series data.
Chapter 2

Data Preparation

2.1 Data Collection

The data collection was conducted in Stockholm from August to October, 2013. Eight commuter cyclists were involved in this organization and their basic information, such as the gender, age, bike type, is summarized in Table 2.1. According to the table, multiple GPS devices were used by the recruited cyclists. Specifically, both Garmin 60CSx and Garmin Edge 500 GPS loggers, respectively, were utilized by three participants, while the rest two cyclists used an Garmin Oregon and an IPhone APP to record their GPS measurements. The participants were required to use GPS devices to record their normal cycling trips. All the devices were regularly brought back to the authors for the data input and the raw data was stored and managed in a SQL database. As Table 2.2 demonstrates, the database includes information about GPS location, altitude, distance and speed. More specifically, all the attributes were measured and logged with a 1-s time interval in the collection process. When this phase of data collection was over, around 100 cycling trips were ultimately collected by the authors and these trips turned out to be completed in different areas of Stockholm. Figure 2.1 then shows four representative trips on the top of an illustration of the Stockholm’s dedicated bike lanes (yellow lines) 1.

While examining the raw data, it was found that different devices documented the measured attributes in different formats, and the missing of information also occurred frequently. Specifically, the information provided by Garmin Edge 500 and Garmin Oregon is

---

1Source: OpenStreetMap, http://www.openstreetmap.org
much more standard and sufficient than that provided by the other two devices, especially the Garmin 60CSx which is a product from nearly 20 years ago. Eventually, the authors decided to abandon all the data collected by Garmin 60CSx and IPhone so that the research can be easier to proceed.

2.2 Preprocessing of Altitude Data

2.2.1 Problem Statement

Altitude is part of the information that can be measured and documented by our GPS devices. However, while checking the altitude data, we found that altitude profiles of many trips were not sufficiently smooth. This is because both Garmin devices (Garmin Edge 500 and Garmin Oregon) measure the altitude through barometric pressure sensors embedded in the devices. With this type of instrument to measure the altitude very frequently (1-s time interval), the Garmin devices can hardly acquire the altitude information consistently. Therefore, many abrupt points can be easily observed in the altitude profiles. Considering that the road gradient is clearly a very significant factor which affects cyclist behavior and this information can only be obtained from the altitude measurements in our research, we decided to process the altitude data so that better quality altitude profiles could serve our later studies.
2.2. PREPROCESSING OF ALTITUDE DATA

Figure 2.1: A demonstration for four representative recorded cycling trips. Yellow lines represent the dedicated cycling tracks in Stockholm. Map data ©Google 2013.

Table 2.2: An illustrative sample of the collected GPS cycling data.

<table>
<thead>
<tr>
<th>Start Time</th>
<th>Time</th>
<th>Latitude</th>
<th>Longitude</th>
<th>Altitude</th>
<th>Distance(m)</th>
<th>Speed(m/s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013/8/30 17:01:29</td>
<td>17:05:50</td>
<td>59.350584</td>
<td>18.069769</td>
<td>19.200001</td>
<td>153.1</td>
<td>4.6</td>
</tr>
<tr>
<td>2013/8/30 17:01:29</td>
<td>17:05:51</td>
<td>59.350539</td>
<td>18.069774</td>
<td>19.200001</td>
<td>156.9</td>
<td>4.6</td>
</tr>
<tr>
<td>2013/8/30 17:01:29</td>
<td>17:05:52</td>
<td>59.350496</td>
<td>18.069779</td>
<td>19.200001</td>
<td>160.9</td>
<td>4.4</td>
</tr>
<tr>
<td>2013/8/30 17:01:29</td>
<td>17:05:53</td>
<td>59.350449</td>
<td>18.069820</td>
<td>19.200001</td>
<td>166.4</td>
<td>4.6</td>
</tr>
<tr>
<td>2013/8/30 17:01:29</td>
<td>17:05:54</td>
<td>59.350401</td>
<td>18.069840</td>
<td>19.200001</td>
<td>171.8</td>
<td>4.8</td>
</tr>
<tr>
<td>2013/8/30 17:01:29</td>
<td>17:05:55</td>
<td>59.350351</td>
<td>18.069855</td>
<td>19.200001</td>
<td>177.5</td>
<td>5.0</td>
</tr>
<tr>
<td>2013/8/30 17:01:29</td>
<td>17:05:56</td>
<td>59.350305</td>
<td>18.069864</td>
<td>19.200001</td>
<td>182.6</td>
<td>5.0</td>
</tr>
<tr>
<td>2013/8/30 17:01:29</td>
<td>17:05:57</td>
<td>59.350254</td>
<td>18.069873</td>
<td>19.200001</td>
<td>189.8</td>
<td>5.2</td>
</tr>
<tr>
<td>2013/8/30 17:01:29</td>
<td>17:05:58</td>
<td>59.350189</td>
<td>18.069886</td>
<td>19.200001</td>
<td>195.6</td>
<td>5.2</td>
</tr>
<tr>
<td>2013/8/30 17:01:29</td>
<td>17:05:59</td>
<td>59.350140</td>
<td>18.069888</td>
<td>19.200001</td>
<td>201.0</td>
<td>5.2</td>
</tr>
<tr>
<td>2013/8/30 17:01:29</td>
<td>17:06:00</td>
<td>59.350094</td>
<td>18.069896</td>
<td>19.200001</td>
<td>206.1</td>
<td>5.2</td>
</tr>
</tbody>
</table>
2.2.2 Methodology

The smoothing of raw altitude profiles was based on the application of robust locally weighted regression. This method was first suggested by Cleveland in 1979 [18] with an intention to enhance the visual information on a scatter plot. Since then this method has been successfully used in many areas to tackle a variety of practical issues.

In the current context of altitude data processing, this methodology can be outlined as follow. Given the time series of the altitude data for a single cycling trip: \( x(t), t = 1, 2, ..., N \), the fitted value at \( x(t) \) is then estimated as a polynomial fit to \( R \) observations (window) in the neighborhood of \( t_0 \). Here the general form of the polynomial function is given as follow

\[
x(t) = f_{t_0}(t, \beta_{t_0}) + \varepsilon_{t_0,t}
\]  

where \( \beta_{t_0} \) denotes the vector of parameters of fitted curve to be estimated and \( \varepsilon_{t_0,t} \) denotes a normally distributed error term. \( f_{t_0}(t, \beta_{t_0}) \), as a whole, represents the fitted position at time \( t \) estimated by a local regression function centered at time \( t_0 \);

Besides, the window size \( R \) is defined as follow

\[
R = f \times N
\]  

where \( f \in (0, 1] \) denotes the parameter which determines the degree of smoothness for the fitted curve and \( N \) denotes the number of measured points.

The weighted least-squares method is then used to estimate the parameters of local function \( f_{t_0}(t, \beta_{t_0}) \) with the \( R \) observations in the window around \( t_0 \). In particular, double weights are assigned to each observation in order to make the local regression robust. The first component of the weight, \( w_{t_0}(t) \), is based on the the normalized time difference, \( d \), between \( t \) and the point of interest \( t_0 \). As \( d \) decreases, the weight of an observation in the window increases, implicating that points close to the target are more significant than those far away from the center. With the first component, the following minimization problem can be formulated, thus yielding the first local polynomial function.

\[
\min_{\beta_{t_0}} |X_{t_0} - f_{t_0}(t, \beta_{t_0})|^T W'_{t_0} [X_{t_0} - f_{t_0}(t, \beta_{t_0})]
\]  

where \( X_{t_0} \) denotes a column vector of \( R \) observations used to estimate the polynomial function centered at \( t_0 \); \( f_{t_0}(t, \beta_{t_0}) \) denotes the corresponding vector of fitted values; \( W'_{t_0} \) denotes a \([N \times N]\) diagonal matrix, with elements corresponding to the first weights of observations which are based on the the normalized time difference.
A column vector of $R$ newly fitted values can be obtained after the first estimation and the residual of each observation is then calculated as

$$e_t = x(t) - \hat{x}(t)$$  \hspace{1cm} (2.4)

Let $s$ be the median of $|e_t|$, so that the second component, the robust weight $\delta_t$, can be calculated as

$$\delta_t = B(e_t/6s)$$  \hspace{1cm} (2.5)

Here $B$ is a bisquare weight function recommended by Cleveland. It is defined as follow

$$B(x) = \begin{cases} 
(1 - x^2)^2 & |x| < 1 \\
0 & \text{otherwise}
\end{cases}$$  \hspace{1cm} (2.6)

In fact, the second component of the weight plays a role in diminishing the effect of outliers. An observation with a large residual is supposed to contain more uncertainty, thus its weight in polynomial fit needs to be decreased accordingly, vice versa. The ultimate weight for each observation, therefore, is a combination of both components, and solving the same minimization problem demonstrated in Equation (2.7) can then obtain the parameters of the desired polynomial function.

$$\min_{\beta_{tr}}[X_{tr} - f_{tr}(t, \beta_{tr}^p)]^T W_{tr}^t [X_{tr} - f_{tr}(t, \beta_{tr}^p)]$$  \hspace{1cm} (2.7)

where $X_{tr}$ denotes a column vector of $R$ observations used to estimate the polynomial function centered at $t_0$; $f_{tr}(t, \beta_{tr}^p)$ denotes the corresponding vector of fitted values; $W_{tr}^t$ denotes a $[N \times N]$ diagonal matrix, with elements corresponding to the first weights of observations which are based on the the normalized time difference.

Eventually, the fitted value at time $t_0$ is calculated by Equation (2.8), and the same process will be just repeated at the next point until the end of the time series.

$$\hat{x}(t_0) = f_{tr}(t, \beta_{tr}^p)$$  \hspace{1cm} (2.8)

### 2.2.3 Practical Results

In practice, a linear function specified by Equation (2.9) was used to fit all the points within the window because such a simple function, as Cleveland [18] suggests, is already capable of providing adequate smoothness.

$$x_{tr}(t) = \beta_{tr,1} t + \beta_{tr,0}$$  \hspace{1cm} (2.9)
As to the weight function, Cleveland et al. [19] pointed out that smooth weight functions lead to smoother estimates. Meanwhile, the selected function should also assign higher weights to observations that are closer to the point of interest. Therefore, a tricube weight function which is also recommended by Cleveland et al. [20] was used in the current application. The tricube function is specified by Equation (2.10) and (2.11), with its shape illustrated in Figure 2.2.

\[ w(t_0, t) = (1 - d(t_0, t)^3)^3 \]  \hspace{1cm} (2.10)
\[ d(t_0, t) = \frac{|t_0 - t|}{D} \]  \hspace{1cm} (2.11)

where \( w(t_0, t) \) denotes the weight assigned to the observation \( x(t) \) in fitting a curve centered at \( t_0 \); \( d(t_0, t) \) denotes the normalized measure of the time difference between \( t_0 \) and \( t \); \( D \) denotes the distance from \( t \) to the nearest point outside the window of \( R \) points. Another key point of implementing this method is to select appropriate degree of smoothness, \( f \). Given the number of measured points \( N \), the window size \( R \) would then solely depend on the magnitude of \( f \). For example, a relatively large \( f \), say, 0.5, would make half of the total points in the window incorporated, hence resulting in fairly smooth and flat curves as displayed in Figure 2.3 (a). On the contrary, a very small \( f \), say, 0.005, would shrink the window size, but represent more details correspondingly (see Figure 2.3 (c)). In the current study, based on the purpose of the processing which is to cancel measurement noise and
Figure 2.3: Smoothed altitude profiles based on three different degrees of smoothness $f$ (0.5, 0.03, 0.005). An $f$ of 0.05 was finally decided for all smoothing cases.
smooth the altitude profiles, the value of $f$ was finally determined to be 0.03 after a number of tests. Figure 2.3 (b) then illustrates a smoothed profile with this setup.

2.3 Preprocessing of GPS Data

2.3.1 Problem Statement

The main function of the Garmin devices is to locate themselves real time through GPS (1-s time interval) and then output useful and consistent information, such as the distance and speed to the users. However, two main problems were found while dealing with the speed and distance profiles. Firstly, missing information can be observed in both kinds of profiles. This is because during the process of data collection, the Garmin devices sometimes failed to log the data for at least one second. Secondly, according to the authors’ detailed examination, the measurement errors in both distance and speed data are not trivial enough to be neglected. Since the aim of this thesis is to model cyclist behavior from a microscopic perspective, better quality data is always desired. In summary, these two specific problems together pose a uneasy challenge to our research.

2.3.2 Methodology

Two approaches were implemented to solve the problems described above. As Figure 2.4 shows, the process starts with the linear interpolation and missing information in speed and distance profiles is thus supplemented. Then based on the first step, the Kalman smoothing algorithm is applied to process both speed and distance profiles.
2.3. PREPROCESSING OF GPS DATA

![Diagram of the ongoing discrete Kalman filter cycle](image)

Figure 2.5: An illustration of the ongoing discrete Kalman filter cycle. The time update projects the current state estimate ahead in time. The measurement update adjusts the projected estimate by an actual measurement at that time.

Introduction of the Kalman Filtering

The Kalman Filtering is a powerful approach to tackle the discrete data linear filtering problem. It was first suggested by Kalman [21] half a century ago. The basic idea of the Kalman Filtering is illustrated by a simple diagram in Figure 2.5. Specifically, the Kalman filter estimates a process using a form of feedback control: the filter estimates the process state at some time and then obtains feedback in the form of (noisy) measurements. In this sense, the Kalman filter comprises two groups of equations: time update equations and measurement update equations. The time update equations are applied to project forward (in time) the current state and error covariance estimates in order to obtain the a priori estimates for the next time step. The measurement update equations, on the other hand, are responsible for the feedback, specifically, for incorporating a new measurement into the a priori estimate to obtain an improved a posteriori estimate.

State space Model

The Kalman Filtering is known for treating non-stationary data processes in an iterative way effectively. It is particularly appropriate for the state-space model form shown as follow

\[
X(t+1) = A(t) \cdot X(t) + G(t) \cdot V(t) \\
Y(t) = H(t) \cdot X(t) + W(t)
\]  

(2.12)
where $\mathbf{X}(t)$ and $\mathbf{Y}(t)$ denotes the state vector and the measurement vector at time $t$, respectively; $A(t)$ and $H(t)$ denotes the state transition matrix and the relation matrix between measurement and state vector at time $t$, respectively; $\mathbf{V}(t)$ denotes the plant noise which is assumed to be a zero-mean Gaussian process; $\mathbf{W}(t)$ denotes the measurement noise which is assumed to be a zero-mean Gaussian process.

Ma et al. in [22] provides a valuable way to process car-following data using the Kalman Filtering, which can be referred in the current study to perform the noise cancellation as well as profile smoothing. The first step, then, is to specify the specific state space model based on the physical state relation revealed by Equation (2.13) and (2.14).

$$s_n(t + 1) = s_n(t) + v_n(t)h + \frac{1}{2}a_n(t)h^2 \quad (2.13)$$

$$v_n(t + 1) = v_n(t) + a_n(t)h \quad (2.14)$$

where $s_n(t)$, $v_n(t)$, and $a_n(t)$ denote the position of object $n$, the speed of object $n$, the acceleration of object $n$ at time $t$, respectively. $h$ denotes the time interval.

In order to complete the state space model, an additional equation concerning the acceleration is required further. Here, according to the order selection approach of autoregressive (AR) models [23], a first-order autoregressive (AR) model described by Equation (2.15) can be applied to represent the acceleration time series. This model represents a random walk process in which the acceleration at the current point of time depends on both the point of time and a random noise term $\theta(t)$. Given a $\phi$ of 1, the time series then become a first-order AR model.

$$a_n(t + 1) = \phi a_n(t) + \theta(t) \quad (2.15)$$

where $\theta(t)$ denotes a white noise following the Gaussian distribution $N(0, \sigma^2)$; $\phi$ denotes a parameter which determines the AR process.

Combining Equation (2.13), Equation (2.14) and (2.15), the state space model can be ultimately summarized as follow

$$\mathbf{X}(t + 1) = A(t) \cdot \mathbf{X}(t) + \mathbf{V}(t) \quad (2.16)$$

$$\mathbf{Y}(t) = H(t) \cdot \mathbf{X}(t) + \mathbf{W}(t) \quad (2.17)$$

where $\mathbf{X}(t) = [s_n(t) \ v_n(t) \ a_n(t)]^T$, $\mathbf{Y}(t) = [\hat{s}_n(t) \ \hat{v}_n(t) \ \hat{a}_n(t)]^T$, $\mathbf{V}(t) = [0 \ 0 \ \theta(t)]^T$, $\mathbf{W}(t) =$
2.3. PREPROCESSING OF GPS DATA

\[ \begin{bmatrix} \varepsilon_s & \varepsilon_v & \varepsilon_a \end{bmatrix}^T, \quad H = I \text{ and} \]

\[
F = \begin{pmatrix}
1 & h & \frac{h^2}{2} \\
0 & 1 & h \\
0 & 0 & \phi
\end{pmatrix}.
\]

The Kalman Smoothing Algorithm

Based on the state space model specified by Equation (2.16), the iterative algorithm of the Kalman Filtering can be then derived as follow

\[
\hat{X}_{t|t-1} = AX_{t-1|t-1} \quad (2.18)
\]

\[
P_{t|t-1} = A_{t-1}P_{t-1|t-1}A_{t-1}^T + Q \quad (2.19)
\]

and

\[
K_t = P_{t|t-1}H^T(HP_{t|t-1}H^T + R)^{-1} \quad (2.20)
\]

\[
\hat{X}_{t|t} = \hat{X}_{t|t-1} + K_t(Y_t - H_t\hat{X}_{t|t-1}) \quad (2.21)
\]

\[
P_{t|t} = (I - K_tH)P_{t|t-1} \quad (2.22)
\]

where \( \hat{X}_{t|t-1} \) and \( \hat{X}_{t|t} \) denote the \textit{a priori} and \textit{a posteriori} estimators, respectively. \( P_{t|t-1} \) and \( P_{t|t} \) denote the \textit{a priori} and \textit{a posteriori} error covariance matrices, respectively. \( Q \) and \( R \) denote the autocovariance matrices of noise processes \( V(t) \) and \( W(t) \), respectively.

Specifically, Equation (2.18) and (2.19) contribute to the time update of the estimator. The former and the latter, respectively, project the state and the error covariance ahead. On the other hand, with the Kalman gain \( K_t \) computed by Equation (2.20), (2.21) and (2.22) can respectively update the estimate and error covariance with the measurement.

Since the smoothing processes were performed offline in the current case, an advantage of knowing the complete observation sequence \( Y_0, \ldots, Y_N \) can be taken to additionally perform a backward filtering. For example, the Rauch-Tung-Striebel smoother, which is an extension of the Kalman Filtering (see [24]), can be implemented in the current case in order to obtain optimal outcomes. It is specified as follow

\[
\begin{align*}
\hat{X}_{t|N} &= \hat{X}_{t|t} + \Omega_t(\hat{X}_{t+1|N} - \hat{X}_{t+1|t}) \quad (2.23) \\
P_{t|N} &= P_{t|t} + \Omega_t(P_{t+1|N} - \hat{X}_{t+1|t})\Omega_t^T \quad (2.24) \\
\Omega_t &= P_{t|t}A_{t}^T \Omega_t^{-1} \quad P_{t+1|t} \quad (2.25)
\end{align*}
\]

Specifically, Equation (2.18) and (2.19) contribute to the time update of the estimator. The former and the latter, respectively, project the state and the error covariance ahead. On the other hand, with the Kalman gain \( K_t \) computed by Equation (2.20), (2.21) and (2.22) can respectively update the estimate and error covariance with the measurement.

Since the smoothing processes were performed offline in the current case, an advantage of knowing the complete observation sequence \( Y_0, \ldots, Y_N \) can be taken to additionally perform a backward filtering. For example, the Rauch-Tung-Striebel smoother, which is an extension of the Kalman Filtering (see [24]), can be implemented in the current case in order to obtain optimal outcomes. It is specified as follow

\[
\begin{align*}
\hat{X}_{t|N} &= \hat{X}_{t|t} + \Omega_t(\hat{X}_{t+1|N} - \hat{X}_{t+1|t}) \quad (2.23) \\
P_{t|N} &= P_{t|t} + \Omega_t(P_{t+1|N} - \hat{X}_{t+1|t})\Omega_t^T \quad (2.24) \\
\Omega_t &= P_{t|t}A_{t}^T \Omega_t^{-1} \quad P_{t+1|t} \quad (2.25)
\end{align*}
\]
CHAPTER 2. DATA PREPARATION

where $\hat{X}_{t|N}$ denotes the estimation of $X_t$ given the data sequence $Y_0, ..., Y_N$; $P_{t|N} = E[(X_t - \hat{X}_{t|N})(X_t - \hat{X}_{t|N})^T]$ denotes the error covariance matrix when the information of the whole data series is used.

In summary, the Kalman smoothing algorithm implemented in this study consists of both forward and backward filtering iterations. As a result, the processed profiles are smoother and more reliable than those which were merely smoothed by using the forward iteration.

2.3.3 Practical Results

Before performing the Kalman smoothing algorithm, missing points in both speed and distance profiles were linearly interpolated in advance. This procedure was critical because experiments show that the Kalman smoother could be greatly affected by abrupt values so that making up these missing points preliminarily markedly helped to enhance the smoothing performance.

When implementing the Kalman smoothing algorithm, multiple parameters were required to be input subjectively, such as $\phi$, $Q$ and $R$. Therefore, it is necessary to specify the parameter setup and discuss why some parameters were particularly used. A $\phi$ of 1 was used for the random walk model (see Equation (2.15)) in order to make the acceleration profile become a first-order ARI process. Meanwhile, since the acceleration measurements were missing, there was no necessity of including it in the measurement model. As a result, $Y(t)$ and $H$ could be simplified as follow

$$Y(t) = [\hat{s}_n(t) \hat{v}_n(t)]^T$$

and

$$H = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

The selection of appropriate noise covariance matrices was another key factor for the filter performance. In the current case, noises of both state and measurement models were assumed to follow Gaussian distribution with a mean of zero. Moreover, the noises of speed and distance in both state and measurement models were assumed to be uncorrelated to each other, thus making both $Q$ and $R$ become diagonal matrices. In addition, although there should be no noise for speed and position in the state update Equation (2.16) since Equation (2.13) and Equation (2.14) are deterministic instead of stochastic, small values (say, 0.1) were given to them in avoidance of the singular matrix. The noise term $\theta(t)$ of the acceleration model in the state update equation is assumed to be constant with a value of 1. Furthermore, Ma et al. in their study [22] mentioned that the performance of the filter
2.3. PREPROCESSING OF GPS DATA

Figure 2.6: An illustrative example of smoothed speed profile and position profile. Plot (a) and (c) show the overall smoothing results for speed and position, respectively, while plot (c) and (d) show some local details.
mainly depends on the ratio of the two covariance matrices and adjusting the ratio can lead to better results. Based on this principle, two matrices $Q$ and $R$ were ultimately set as follow:

$$Q = \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0.1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad R = \begin{pmatrix} 5 & 0 \\ 0 & 15 \end{pmatrix}.$$ 

An illustrative example of smoothed speed profile and position profile were demonstrated in Figure 2.6. Besides plot (a) and (c) which show the overall smoothing results for speed and position, respectively, two additional plots (b) and (d) were also included in order to present some local details. According to the first two plots (a) and (b), the speed profile was significantly changed and became smoother accordingly. However, the last two plots indicate that this implementation did not have a remarkable influence on the position profile. Only slight variation can be observed even in the zoom-in plot (Figure 2.6 (d)).

### 2.4 Acceleration Profile Estimation

Since the GPS devices were not able to measure cyclists’ instantaneous acceleration rate, it has to be estimated through other measurements. In the current case, the acceleration rate was finally derived from the speed profile with a time interval of one second, and the method used was suggested by Wang et al. [25] in a study regarding the GPS data of car drivers. According to the method, the acceleration rate at the $n$th second is estimated by averaging that of the previous second and of the following second. Given the time series of measurements of speed for a cyclist: $v_1, v_2, ..., v_n$, the acceleration rate at a certain second is calculated by Equation (2.26) and an example of derived acceleration profile is shown in Figure 2.7.

$$a(t) = \frac{v(t + 1) - v(t - 1)}{2}, \quad 1 < t < n \quad (2.26)$$

where $a(t)$ denotes the estimated acceleration rate at time $t$ in meters per second squared; $v(t)$ denotes the speed at time $t$ in meters per second. Particularly, $a(1)$ and $a(n)$ are set to be 0 by default.
Figure 2.7: An illustration of a processed speed profile and the acceleration profile derived from it.

2.5 Gradient Profile Estimation

The diagram in Figure 2.8 demonstrates the algorithm about how the gradient at time $t(t > 1)$ was computed. It has to be pointed out that in the current case, the gradient at time $t$ is used to reflect the change in the terrain between time $t-1$ and $t$. In order to make up the inaccuracy caused by the GPS devices, a minimum 10-meter constraint was added, which means the gradient could only be calculated when the distance had been larger than 10 meters. In addition, the gradient at the first point of time was set to be 0 by default.

2.6 Summary

This chapter comprises the presentation of data collection and data preprocessing performed in the study. By letting individual cyclists use portable GPS devices, naturalistic data which
record the cycling performance were gathered. The raw measurements include the position, altitude, speed and so on with time interval of one second. As to data preprocessing, two parts are included. On the one hand, data-smoothing approaches, such as the locally weight regression and the Kalman filtering, were implemented to reduce measurement errors. On the other hand, additional attributes, such as the instantaneous acceleration rate and road gradient, were derived from direct measurements.

Figure 2.8: A demonstration for the gradient calculation algorithm.
Chapter 3

Model Development

3.1 The Cycling Regime

In the current study, a cyclist is assumed to only have three different behavioral options while cycling. These options consist of the acceleration behavior, deceleration behavior, and cruising behavior. Based on this particular classification, the cyclist behavior can be simply explained by a regime that during the cycling process, a cyclist is always endeavoring to achieve and maintain a desired speed which varies real-time depending on multiple factors, such as the age and gender of the individual (internal factors), and the road grade and weather (external factors). The processes during which the cyclist is maintaining her present desired speed then corresponds to the cruising behavior, while the processes during which she tries to achieve her desired speed refers to either the acceleration behavior or the deceleration behavior. Further, this regime can be summarized by Equation (3.1).

\[
a_n(t) = \begin{cases} 
a_n^{acc}(t) & \text{if } v_n(t) < v_n^d \\
0 & \text{if } v_n(t) > v_n^d \\
\end{cases} \quad (3.1)
\]

where \(a_n(t)\) denotes an individual cyclist’s acceleration rate at time \(t\); \(a_n^{acc}(t)\) and \(a_n^{dec}(t)\), respectively, denote the acceleration rate and deceleration rate at time \(t\); \(v_n^d\) denotes the cyclist’s desired speed at time \(t\) which depends on the cyclist’s demographics and environmental factors, such as the road gradient.
Table 3.1: An illustration about how the acceleration and deceleration profiles in the current study are selected stepwise.

**Step 1:**
Distinguishing the continuous speeding-up and slowing-down data points and marking each cluster as a candidate profile.

**Step 2:**
Ruling out candidate profiles that do not meet the preliminary criteria listed in Table 3.2.

**Step 3:**
Computing the proportional variation in speed using Equation (3.2) for all the remained candidate profiles and further eliminate insignificant profiles.

### 3.2 The Cycling Performance

#### 3.2.1 Profile Selection

**The Acceleration and Deceleration Profiles**

With the suggested cycling regime, selecting appropriate sub-datasets which involves cyclist acceleration behavior or deceleration behavior becomes critical for further research. In the current case, these sub-datasets are termed as acceleration or deceleration profiles. Specifically, an acceleration profile refers to a part of the trip when cyclists accelerate from a low initial speed to a high final speed, while a deceleration profile means a part of the trip when cyclists decelerate from a high initial speed to a low final speed.

As Table 3.1 illustrates, it takes three steps to obtain the desired information about acceleration and deceleration from a large amount of GPS data. As the first procedure, finding the cycling processes during which cyclists were continuously speeding up or slowing down provides candidate profiles for the following filtering steps. Then five rigorous criteria shown in Table 3.2 which were made based on realistic observations and tests, were executed to rule out insignificant profiles. As a result, 5331 acceleration profiles and 5373 deceleration profiles remained, respectively. Finally, on the basis of former two steps, a new index was added in order to further eliminate even less significant profiles. Specifically, this index is termed as the proportional variation in speed and it reflects the degree of change in speed during an acceleration or deceleration process. It is calculated as
3.2. THE CYCLING PERFORMANCE

Table 3.2: Preliminary selection criteria for acceleration and deceleration profiles

<table>
<thead>
<tr>
<th>Criteria</th>
<th>An acceleration profile \ A deceleration profile;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Maximum acceleration $\in (0, 2]$ (m/s$^2$) \ Minimum deceleration $\in [-2,0)$(m/s$^2$);</td>
</tr>
<tr>
<td>2</td>
<td>Final speed $\in (0,15]$ (m/s) \ Initial speed $\in (0,15]$ (m/s$^2$);</td>
</tr>
<tr>
<td>3</td>
<td>The duration of entire process $\in [5,20]$ (sec);</td>
</tr>
<tr>
<td>4</td>
<td>The moving distance during the entire process $\in [5, +\infty)$ (m);</td>
</tr>
<tr>
<td>5</td>
<td>The maximum and minimum gradient during the process $\in [-7, 7]$ (%).</td>
</tr>
</tbody>
</table>

\[
\eta = \frac{|\Delta V|}{v_f} \times 100\% = \frac{|v_f - v_i|}{v_f} \times 100\% \tag{3.2}
\]

where $\eta$ denotes the proportional variation in speed; $v_i$ and $v_f$, respectively, denote the initial speed and final speed of an acceleration (deceleration) process.

The index was calculated for all remained profiles, with the distributions of both acceleration’s and deceleration’s results plotted in Figure 3.1. According to the diagrams, most profiles merely exhibit minor increases or decreases in speed (lower than 25 percent) and the frequency declines markedly as $\eta$ gradually grows. At the end, however, both diagrams witness a soar in the frequency, showing that there are nearly 200 profiles for both cases in which the cycling speed varied by 100 percent. Hence, these are profiles which document cyclists’ complete starts of stops.

Given the information provided by the distributions shown in Figure 3.1, a critical value regarding the proportional increase, or decrease, in speed is then required to further rule out insignificant acceleration, or deceleration, profiles. In Figure 3.2, four line charts illustrating how the number of acceleration and deceleration profiles as well as the number of data points involved in these profiles decrease as the critical value increases are presented. Based on these charts and a perception on cyclists’ characteristics, 50\% was eventually determined as the critical value for the current case, which means only those profiles of which proportional change in speed is greater than 50\% could be labeled an acceleration or deceleration profile. As a result, 665 acceleration profiles and 673 deceleration profiles were selected, respectively.
CHAPTER 3. MODEL DEVELOPMENT

Figure 3.1: Distributions of the proportional variation in speed for both acceleration and deceleration cases.

Figure 3.2: A demonstration about how the number of acceleration and deceleration processes and points included in those processes varies as the threshold increases.
### 3.2. THE CYCLING PERFORMANCE

#### Table 3.3: Selection criteria for the cruising profiles.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>A point eligible for the cruising profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Not included in the acceleration or deceleration profiles with a value of ( \eta ) larger than 20%;</td>
</tr>
<tr>
<td>2</td>
<td>Speed ( \in [2, 15] \text{ m/s}; )</td>
</tr>
<tr>
<td>3</td>
<td>Acceleration rate ( \in [-0.2, 0.2] \text{ m/s}^2 );</td>
</tr>
<tr>
<td>4</td>
<td>Gradient ( \in [-7, +7] % ).</td>
</tr>
</tbody>
</table>

#### The Cruising Profiles

A dataset for the cruising behavior was set up with the data adopted by the acceleration and deceleration profiles excluded. As Table 3.3 shows, four criteria were used to select appropriate cruising profiles. The first criterion implies that those cycling processes with small fluctuations (lower than 20%) were also regarded as normal cruising performance. The other three criteria were formulated based on the perception on ordinary cycling activities. As a result, a dataset containing 48579 tuples were created.

#### 3.2.2 Profile Analysis

**The Acceleration and Deceleration Performance**

Further examination of selected acceleration and deceleration profiles led a finding that the speed-time curve typically had an S shape as shown in Figure 3.3. Accordingly, the acceleration-time curves exhibited a U shape, starting and ending at zero acceleration rate. With more cases presented in Appendix A, however, it has to be pointed out that due to the large variation in data, the acceleration and speed profiles are not often as smooth as indicated by data points in Figure 3.3 which indeed demonstrates very typical profiles. An important feature of Figure 3.3 is that it emphasizes the physical requirements for the model of zero acceleration, \( a = 0 \), and zero jerk, \( da/dt = 0 \), at the start and end of both acceleration and deceleration cases (time 0 and time \( t \)). It will then be a critical factor for the model development.

Four key attributes were then extracted or calculated for each profile which is also termed as a process in the following context. They are

---

[Note: The table and the text provide a structured overview of the selection criteria for cruising profiles, followed by an in-depth analysis of accelerated and decelerated profiles, emphasizing the importance of smoothness and physical requirements.]
Figure 3.3: Typical acceleration (left) and deceleration (right) profiles, as well as their corresponding speed profiles.

1) The initial speed: the first value of second-by-second speed during a process;

2) The final speed: the last value of second-by-second speed during a process;

3) The average acceleration (deceleration) rate: the mean value of second-by-second acceleration (deceleration) rates during a process;

4) The maximum acceleration (deceleration) rate: the peak value of second-by-second acceleration (deceleration) rates during a process.

The results were summarized and presented through histograms shown in Figure 3.4 and Figure 3.5, with the former and the latter illustrating the acceleration case and deceleration case, respectively. Four histograms in each figure, respectively, show the distribution of the initial speed, the final speed, the average acceleration rate, and the maximum acceleration rate for selected acceleration and deceleration profiles. These plots intuitively reveal significant features of cyclists’ acceleration and deceleration performance. For instance, Diagram
3.2. THE CYCLING PERFORMANCE

Figure 3.4: Key attributes’ distributions for acceleration profiles

Figure 3.5: Key attributes’ distributions for deceleration profiles.
(a) in Figure 3.4 shows that cyclists were much less likely to speed up when they had been already cycling at a relatively high speed (greater than 4 m/s), while complete starts (in which the initial speed is 0) account for a large part of acceleration cases. Moreover, according to Diagram (b) in Figure 3.4, most acceleration processes ended up at a speed level of 5 m/s, although there are also plenty of cases showing that cyclists did not stop speeding up until they reached around 10 m/s. As to the other attributes, Diagram (c) and (d) in Figure 3.4 imply a fact that the average acceleration rate is always smaller than the maximum acceleration rate. The average acceleration can hardly exceed 1 m/s² but the maximum one during a process is even able to peak at 2 m/s². Meanwhile, for the acceleration case’s counterpart-deceleration, the plots in Figure 3.5 turn out to resemble those in Figure 3.4 except that the distribution of the initial speed in deceleration case is similar to that of the final speed in acceleration case, vice versa.

In order to have a deeper understanding of relations among all variables associated with an acceleration or deceleration process, correlation and regression analysis were conducted further. Correlation analysis is always used as an effective method to examine whether there is a linear relation between two variables in the time domain. As the main evidence, the index-correlation coefficient-accounts for the general strength of the linear relationship between two variables (namely, two data vectors). The index is defined as follow

$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$  \hspace{1cm} (3.3)

where $\mu_X$ denotes the mean of variable $X$ and $\sigma_X$ denotes the standard deviation of variable $X$, the same for variable $Y$. However, it is the sample correlation coefficient that was normally used in practice. It is defined as

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{s_X} \right) \left( \frac{Y_i - \bar{Y}}{s_Y} \right)$$  \hspace{1cm} (3.4)

where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$, $s_X = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$.

The calculation results of correlation coefficients are presented in Table 3.4 and Table 3.5 for the acceleration case and deceleration case, respectively. It is notable that both the average acceleration $a_{avg}$ and maximum acceleration $a_{max}$ show significant linear relationship with $\Delta v$ which is a variable reflecting the change in speed during a process (the same for the deceleration case). This finding then led regression analyses among three related variables ($\Delta v, a_{max}$ and $a_{avg}$, $\Delta v, d_{min}$ and $d_{avg}$). As it can be expected, the linear relationship
3.2. THE CYCLING PERFORMANCE

Table 3.4: The calculation results of correlation coefficients for the acceleration case.

<table>
<thead>
<tr>
<th>Initial speed $v_i$</th>
<th>Final speed $v_f$</th>
<th>Speed increase $\Delta V$</th>
<th>Maximum acceleration $a_{\text{max}}$</th>
<th>Average acceleration $a_{\text{avg}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_i$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$v_f$</td>
<td>0.66</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta v$</td>
<td>0</td>
<td>0.75</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$a_{\text{max}}$</td>
<td>-0.01</td>
<td>0.5</td>
<td>0.67</td>
<td>1</td>
</tr>
<tr>
<td>$a_{\text{avg}}$</td>
<td>-0.01</td>
<td>0.46</td>
<td>0.62</td>
<td>0.93</td>
</tr>
</tbody>
</table>

Table 3.5: The calculation results of correlation coefficients for the deceleration case.

<table>
<thead>
<tr>
<th>Initial speed $v_i$</th>
<th>Final speed $v_f$</th>
<th>Speed decrease $\Delta V$</th>
<th>Maximum deceleration $d_{\text{max}}$</th>
<th>Average deceleration $d_{\text{avg}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$v_i$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$v_f$</td>
<td>0.68</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\Delta v$</td>
<td>-0.76</td>
<td>-0.04</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$d_{\text{max}}$</td>
<td>-0.65</td>
<td>-0.23</td>
<td>0.69</td>
<td>1</td>
</tr>
<tr>
<td>$d_{\text{avg}}$</td>
<td>-0.59</td>
<td>-0.19</td>
<td>0.63</td>
<td>0.93</td>
</tr>
</tbody>
</table>

between $\Delta v$ and $a_{\text{max}}$, and $\Delta v$ and $a_{\text{avg}}$ are revealed by the scatter plots shown in Figure 3.6 (a) and Figure 3.6 (b), respectively. However, both plots demonstrate heteroscedasticity at the same time, with the oscillation of the maximum acceleration rate and average acceleration rate increasing as $\Delta v$ grows (the same for deceleration cases shown in plot (a) and (b) in Figure 3.7). Therefore, all variables were further transformed by computing its natural logarithm before the regression analysis. Consequently, the heteroscedasticity was largely ruled out as displayed by scatter plots in Figure 3.6 (c) and (d). More importantly, the values of R square, in the meanwhile, increase to 0.48 and 0.41, respectively as well, making it reasonable to predict both the maximum acceleration and average acceleration of an acceleration process through $\Delta v$ (the same for the deceleration case).
Figure 3.6: Plots of regression analysis for the acceleration case.

Figure 3.7: Plots of regression analysis for the deceleration case.
3.2. THE CYCLING PERFORMANCE

Table 3.6: Linear regression models for the cruising speed.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>t-statistic</td>
</tr>
<tr>
<td>Constant</td>
<td>6.132</td>
<td>1071.38</td>
</tr>
<tr>
<td>Gradient (%)</td>
<td>-0.359</td>
<td>-121.75</td>
</tr>
<tr>
<td>Downhill gradient (%)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Uphill gradient (%)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Adjusted R-squared</td>
<td>0.234</td>
<td></td>
</tr>
<tr>
<td>Observations</td>
<td>48597</td>
<td></td>
</tr>
</tbody>
</table>

The Cruising Performance

Linear regression models were constructed to investigate the cruising performance. Since the gradient was found to be significantly correlated to the selected cruising speed (with the correlation coefficient equal to -0.4835), it was used as the major independent variable for the regression. Other person type variables, including the age, gender and bicycle type, were not found to have influences on the cruising performance significantly, which may be attributed to the small sample size.

The estimation results of two regression models are shown in Table 3.6, with model 2 differentiating the uphill gradient and downhill gradient. Both models suggest that when the gradient is zero, namely when the cyclists are moving forward on the flat, the mean speed of them is around 6.1 m/s², which is very close to Parkin et al.’s finding in [14] (6.01 m/s²). The slight difference could be caused by the present case’s specific data selection as Parkin et al. did not differentiate the acceleration and deceleration data with the cruising data. In addition, although both models yield the same value of adjusted R-squared, different coefficients for downhill and uphill gradient in model 2 implies a fact that cyclists’ sensitivity to the gradient is different in two cases. Specifically, the cycling speed drops more rapidly in the uphill case than in the downhill case because the coefficient of uphill gradient (-0.371) is smaller than that of downhill gradient (-0.339). This result is coherent with Parkin et al.’s explanation in [14] that cyclists will break at steeper gradients and will not take advantage of the additional potential energy they are gaining.
3.3 The Acceleration Model

3.3.1 The Polynomial Model

Although acceleration profiles vary widely among individual cyclists and are also dependent on a number of external factors, such as the bicycle type and weather conditions, they share the same acceleration and speed forms as shown in Figure 3.3 (a) and (b). In fact, this feature is not particular to cycling profiles. Similar S-shape speed-time profiles and U-shape acceleration-time profiles can be observed in driving acceleration profiles as well. One of the classical studies which endeavored to investigate drivers’ acceleration and deceleration profile was conducted by Akçelik et al. in 1987 [26]. In the study, Akçelik et al. used three different models, including a two-term sinusoidal, a three-term sinusoidal, and a polynomial model to estimate acceleration profiles and further compare the performances of different models. As a result, the polynomial model was found to be able to provide best estimations. Therefore, this polynomial model was also referred by the author and became the basis of the model development.

The general function of the polynomial model is specified by Equation (3.5).

\[ y = ax^n(1 - x^m)^2 \]  

where parameter \( a \) scales the final output of the function, while parameter \( n \) and \( m \) together
3.3. THE ACCELERATION MODEL

determine the specific shape of the curve. Diagrams in Figure 3.6 then illustrate
with a fixed $a$ of 1 and varying values of $n$ and $m$ in Figure , it can be known that both
parameters turn out to play roles in shifting the peak point between 0 and 1. Furthermore,
it is also worth mentioning that an advantage of this polynomial model is that it meets the
following requirements of acceleration and speed profiles:

1) zero acceleration at the start and end of the acceleration;

2) zero jerk ($dy/dx = 0$) at the start and end of the acceleration.

3.3.2 The Acceleration Model

The main idea of developing the acceleration model was to model cyclists’ acceleration
profiles referring to the polynomial function specified by Equation (3.5). In this sense, the
general form of the acceleration model was proposed as follow

$$a_{acc}(t) = \alpha \cdot (\Delta V_k)^2 \cdot \theta_k(t)^p \cdot (1 - \theta_k(t))^q + \epsilon_{acc}(t)$$  (3.6)

where $a_{acc}(t)$ denotes the acceleration rate at time $t$; $\Delta V_k$ denotes the speed difference
between the cyclist’s present speed and her desired speed during the $k$th acceleration process;
$\alpha$, $\beta$, $p$ and $q$ denote parameters to be determined; $\theta_k(t)$ denotes a formulated independent
variable, which is defined as follow

$$\theta_k(t) = \frac{v(t) - v_{k,i}}{v_{k,f} - v_{k,i}} \cdot \frac{v(t) - v_{k,i}}{\Delta V_k}$$  (3.7)

where $v(t)$ denotes the cycling speed at time $t$; $v_{k,i}$ and $v_{k,f}$, respectively, denote the initial
and final speed of the $k$th acceleration process.

$\epsilon_{acc}(t)$ denotes the random term associated with the acceleration rate at time $t$. This
random term captures the effect of omitted variables and is assumed to follow a zero-mean
normal distribution. It is also assumed to be independent for different cyclists and different
acceleration processes. Further, acceleration decisions within each acceleration processes
are assumed to be uncorrelated, hence resulting in implications shown as follow

$$\epsilon_{acc}(t) \sim N(0, \sigma^2)$$  (3.8)

$$\text{cov}(\epsilon_{acc}(t), \epsilon_{acc}(t')) = \begin{cases} \sigma^2 & \text{if } t = t', n = n' \\ 0 & \text{otherwise} \end{cases}$$  (3.9)
3.3.3 Likelihood Function Formulation

Given Equation (3.6) and the observed speed data as well as derived acceleration data, the model parameters $\alpha, \beta, p$ and $q$ can be estimated using maximum likelihood estimator (MLE). The probability density function of the acceleration model is hence given by

$$f(a_{\text{acc}}(t)) = \frac{1}{\sigma \phi} \left( \frac{a_{\text{acc}}(t) - \alpha \cdot (\Delta V_k)^\beta \cdot \theta_k(t)^p (1 - \theta_k(t)^q)^2}{\sigma} \right)$$ (3.10)

Then the likelihood function for $K$ acceleration processes can be formulated as follow

$$\mathcal{L}(\alpha, \beta, p, q | a_{\text{acc}}(t)) = \prod_{k=1}^{K} \prod_{t=1}^{T_k} f(a_{\text{acc}}^k(t))$$ (3.11)

Finally, the log-likelihood function for all observations is given by

$$\mathcal{L}L = \sum_{k=1}^{K} \sum_{t=1}^{T_k} \ln[f(a_{\text{acc}}^k(t))] = -\frac{N_o}{2} \ln(2\pi) - N_o \ln \sigma - \frac{1}{2\sigma^2} \left[ \sum_{k=1}^{K} \sum_{t=1}^{T_k} \left( a_{\text{acc}}^k(t) - \alpha \cdot (\Delta V_k)^\beta \cdot \theta_k(t)^p (1 - \theta_k(t)^q)^2 \right)^2 \right]$$ (3.12)

where $N_o$ denotes the number of the whole observations and is given by

$$N_o = \sum_{i=1}^{K} T_i$$ (3.13)

Maximizing the likelihood function would provide the MLE estimate of the model parameters.

3.3.4 Model Specification

Three model specifications are presented in this section. Through comparing different combinations of the parameters, the best fitted one can be selected.

**Model 1**

The first model specification only keeps the very basic but indispensable parameters: $\alpha$ and $q$. $\beta$ and $p$ are omitted. As a results, this one can be used as a baseline for other model specifications.

$$a_{\text{acc}}(t) = \alpha \cdot \Delta V_k \cdot \theta_k(t) (1 - \theta_k(t)^q)^2 + \varepsilon_{\text{acc}}(t)$$ (3.14)
3.3. THE ACCELERATION MODEL

Model 2

The second specification lets parameter $\beta$ enter the model so that $\Delta V$’s effect on the final acceleration can be reflected through $\beta$.

\[
a^{acc}(t) = \alpha \cdot (\Delta V_k)^{\beta} \cdot \theta_k(t) (1 - \theta_k(t)^p)^2 + \varepsilon^{acc}(t)
\]  

(3.15)

Model 3

The third model specification retains all the parameters that show up in the general form. The purpose of this specification is to examine parameter $p$’s influence on the shape of the model curve and to see whether this parameter is necessary for the developed model.

\[
a^{acc}(t) = \alpha \cdot (\Delta V_k)^{\beta} \cdot \theta_k(t)^p (1 - \theta_k(t)^p)^2 + \varepsilon^{acc}(t)
\]  

(3.16)

Model 4 & Model 5

Model 4 (Equation (3.17)) and Model 5 (Equation (3.18)) are extended from Model 2 and Model 3, respectively, with the general stimulus $\beta$ replaced by a linear function which includes two explanatory variables: Gender and Bike type. The purpose of this replacement is to figure out whether person type variables have an influence on the cyclists’ acceleration behavior. Since $\Delta v$ varies in every acceleration process, it can be reasonable to speculate that the influence can be captured through this variable instead of others.

\[
a^{acc}(t) = \alpha \cdot (\Delta V_k)^{\beta_g X_g + \beta_{rb} X_{rb}} \cdot \theta_k(t) (1 - \theta_k(t)^p)^2 + \varepsilon^{acc}(t)
\]  

(3.17)

\[
a^{acc}(t) = \alpha \cdot (\Delta V_k)^{\beta_g X_g + \beta_{rb} X_{rb}} \cdot \theta_k(t)^p (1 - \theta_k(t)^p)^2 + \varepsilon^{acc}(t)
\]  

(3.18)

where $X_g$ and $X_{rb}$, respectively, denote the explanatory variables of cyclists’ gender and bike type. They enter the model as dummy variables as Equation show. $\beta_g$ and $\beta_{rb}$, respectively, denote their corresponding parameters.

\[
\text{Gender} = \begin{cases} 
1 & \text{male} \\
0 & \text{female} 
\end{cases}
\]  

(3.19)

\[
\text{Bike type} = \begin{cases} 
1 & \text{Racing bike} \\
0 & \text{others} 
\end{cases}
\]  

(3.20)
Table 3.7: Estimation results for the acceleration models (t-statistics in parenthesis).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \alpha )</td>
<td>0.6456 (44.69)</td>
<td>1.0948 (20.49)</td>
<td>0.2353 (24.38)</td>
<td>2.2319 (30.50)</td>
<td>0.5398 (38.15)</td>
</tr>
<tr>
<td>( q )</td>
<td>1.7520 (43.54)</td>
<td>1.7693 (44.38)</td>
<td>10.0763 (33.27)</td>
<td>1.7867 (41.91)</td>
<td>9.5296 (31.24)</td>
</tr>
<tr>
<td>( \beta )</td>
<td>–</td>
<td>0.6723 (24.45)</td>
<td>0.7023 (29.24)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>( p )</td>
<td>–</td>
<td>–</td>
<td>0.1661 (17.94)</td>
<td>–</td>
<td>0.1718 (16.88)</td>
</tr>
<tr>
<td>Gender ( (\beta_g) )</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.1711 (10.94)</td>
<td>0.1642 (12.05)</td>
</tr>
<tr>
<td>Bike type ( (\beta_{bt}) )</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.1352 (8.60)</td>
<td>0.1309 (9.53)</td>
</tr>
<tr>
<td>( \sigma ) ( (\text{m/s}^2) )</td>
<td>0.2774 (109.8)</td>
<td>0.2743 (109.8)</td>
<td>0.2447 (109.8)</td>
<td>0.2858 (109.8)</td>
<td>0.2592 (109.8)</td>
</tr>
<tr>
<td>Log likelihood (0)</td>
<td>–7581.2</td>
<td>–7734.0</td>
<td>–9611.7</td>
<td>–7195.6</td>
<td>–8588.7</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>–823.4</td>
<td>–756.4</td>
<td>–67.2</td>
<td>–1003.0</td>
<td>–415.8</td>
</tr>
<tr>
<td>Observations</td>
<td>6028</td>
<td>6028</td>
<td>6028</td>
<td>6028</td>
<td>6028</td>
</tr>
</tbody>
</table>

3.3.5 Estimation Results

For the sake of later model validation (cross validation), only 80% profiles of each individual cyclist were randomly selected and used for the model estimation. Specifically, 6028 observations remained for this procedure. The maximum likelihood estimation was performed analytically using the code programmed by the author in MATLAB R2013a [27]. Multiple functions provided by MATLAB, such as the `nlinfit`, `mle` and `acov`, were implemented. The detailed MATLAB code can be reviewed in Appendix B.

The estimation results for all acceleration models are summarized in Table 3.7. As it shows, Model 1 is illustrated in the first column as the most fundamental case where only very basic parameters \( \alpha \) and \( q \) stay in the model. After \( \beta \) enters the model as Model 2 demonstrates, the magnitude of \( \alpha \) then immediately increases by around 0.4. This is because the parameter \( \beta \), which is smaller than 1 shrinks the effect of \( \Delta V \) on the final value of acceleration so that the parameter \( \alpha \) has to accordingly increment to compensate for the shortage in the estimate. In the meanwhile, Model 2 also sees an increase of 67 in the log likelihood, indicating that Model 2 provides better estimation results.

The estimation result of Model 3 portrays some different features. Notably, the value of \( \alpha \) plunges to 0.2353 as the parameter \( p \) is ultimately included. At the same time, the value of \( q \) roars to 10.0763 while the change in the value of \( \beta \) is not remarkable. Furthermore, Model 3 outcompetes the other models significantly in terms of the Log likelihood (–67.2).
3.3. THE ACCELERATION MODEL

The last two columns of Table 3.7 illustrate the estimation results of the fourth and fifth model specifications. Two person-type explanatory variables enter Model 4 and Model 5, partly accounting for the variation in the acceleration. It is notable that the estimates of some old variables in these two models, such as $p$ and $q$, do not show significant changes compared with their counterparts (Model 2 and Model 3), while the value of $\alpha$ in both extended models nearly doubles. Both $\beta_g$ and $\beta_{bt}$ in Model 4 and Model 5 turn out to be positive and statistically significant, implying that male cyclists and cyclists with a racing bike accelerate harder and more fiercely. However, it can be also seen that both extensions see a great decrease in the Log likelihood ($-1003.0$ and $-415.8$), meaning that the extended models do not provide desirable estimates. This can be attributed to the limitation of the data's diversity. Other factors, such as cyclists’ age and physical condition, may also have critical effects on cyclists’ acceleration performance. In the future work, this issue can be examined again.

3.3.6 Model Validation

Model validation was performed based on the speed profiles. Specifically, the speed profiles for the validation were derived by using the developed acceleration models. Then the observed speed profiles were compared with the derived ones depending on multiple goodness-of-fit measures, including the root mean square error (RMSE) and Theil’s inequality coefficient (U) which quantify the overall error of the validation; the mean error (ME) and mean absolute percentage error (MAPE) which reflect the existence of systematic under- or over- prediction by the developed models. The equations are shown as follow

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (Y_{pn} - Y_{on})^2} \quad (3.21)
\]

\[
\text{MAPE} = \frac{1}{N} \sum_{n=1}^{N} \left| \frac{Y_{pn} - Y_{on}}{Y_{pn}} \right| \quad (3.22)
\]

\[
\text{ME} = \frac{1}{N} \sum_{n=1}^{N} (Y_{pn} - Y_{on}) \quad (3.23)
\]

\[
U = \frac{\sqrt{\frac{1}{N} \sum_{n=1}^{N} (Y_{pn} - Y_{on})^2}}{\sqrt{\frac{1}{N} \sum_{n=1}^{N} (Y_{pn})^2}} \quad (3.24)
\]

where $Y_{pn}$ and $Y_{on}$, respectively, denote the nth predicted and observed measurements.
Table 3.8: Statistics for the speed comparison for the acceleration case.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE (m/s)</td>
<td>1.2</td>
<td>1.2</td>
<td>0.8</td>
<td>1.3</td>
<td>0.9</td>
</tr>
<tr>
<td>MAPE (%)</td>
<td>22.4</td>
<td>21.8</td>
<td>14.9</td>
<td>22.2</td>
<td>16.1</td>
</tr>
<tr>
<td>ME (m/s)</td>
<td>−0.73</td>
<td>−0.70</td>
<td>−0.11</td>
<td>−0.73</td>
<td>−0.13</td>
</tr>
<tr>
<td>U (fraction)</td>
<td>0.0043</td>
<td>0.0044</td>
<td>0.0027</td>
<td>0.0049</td>
<td>0.0032</td>
</tr>
</tbody>
</table>

Table 3.8 summarizes the results for all goodness-of-fit measures. According to them, Model 3 yields the best outcome with regards to all statistics, while Model 1 and Model 2 somehow result in quite similar ones. Besides the specific statistical values, an example of validation case can be also viewed in Figure 3.9. The upper plot in the figure clearly shows that the speed curve resulting from Model 3 resembles the observed one most. As for Model 4 and Model 5, both of them merely see slight reduction in all statistics compared with their counterparts (Model 2 and Model 3). The plots in Figure 3.10 also show that only slight differences in both acceleration profile and speed profile between the original models and the extended model can be observed.

3.4 The Deceleration Case

As plots (b) and (d) in Figure 3.3 portray, the cyclists’ deceleration profile is almost the mirror image of their acceleration profile. Therefore, the same polynomial function was utilized to model the deceleration behavior, with Table 3.9 and Table 3.10 presenting the estimation results and validation statistics, respectively.

Differing from the acceleration case, the deceleration one does not yield sound validation statistics. For example, all models except Model 3 and its extension Model 5 result in MAPE which is larger than 100%. The great underestimate by Model 1 and Model 2 can be further examined in Figure 3.11 which illustrates predicted speed profiles and deceleration profiles by three models. As the figure shows, Model 1 and Model 2 basically fail to describe the deceleration behavior over the given period, while Model 3 still functions quite properly, which corresponds to Model 3’s validation statistics.
3.4. THE DECELERATION CASE

Figure 3.9: An example of predicted speed profiles and acceleration profiles by Acceleration Model 1, Model 2 and Model 3.

Figure 3.10: Comparisons between Model 2 and Model 4, and between Model 3 and Model 5.
Table 3.9: Estimation results for the deceleration models (t-statistics in parenthesis).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$-0.35 (-81.2)$</td>
<td>$-0.56 (-25.0)$</td>
<td>$-0.26 (-29.5)$</td>
<td>$-1.06 (-42.4)$</td>
<td>$-0.52 (-48.3)$</td>
</tr>
<tr>
<td>$q$</td>
<td>4.55 (50.2)</td>
<td>4.57 (50.9)</td>
<td>14.2 (36.7)</td>
<td>4.61 (48.5)</td>
<td>13.83 (34.6)</td>
</tr>
<tr>
<td>$\beta$</td>
<td>$-0.71 (29.5)$</td>
<td>$0.71 (35.4)$</td>
<td>$-0.32 (36.4)$</td>
<td>$0.20 (15.3)$</td>
<td>$0.33 (34.4)$</td>
</tr>
<tr>
<td>$p$</td>
<td>$-0.23$ (29.5)</td>
<td>$-0.26$ (35.4)</td>
<td>$-0.23$ (22.9)</td>
<td>$-0.28$ (25.1)</td>
<td>$-0.28$ (25.1)</td>
</tr>
<tr>
<td>Gender ($\beta_g$)</td>
<td>$-0.20 (15.3)$</td>
<td>$0.20 (15.3)$</td>
<td>$0.20 (15.3)$</td>
<td>$0.20 (15.3)$</td>
<td>$0.20 (15.3)$</td>
</tr>
<tr>
<td>Bike type ($\beta_{bt}$)</td>
<td>$-0.20 (15.3)$</td>
<td>$0.20 (15.3)$</td>
<td>$0.20 (15.3)$</td>
<td>$0.20 (15.3)$</td>
<td>$0.20 (15.3)$</td>
</tr>
<tr>
<td>$\sigma$ ($m/s^2$)</td>
<td>0.25 (110.5)</td>
<td>0.24 (110.5)</td>
<td>0.21 (110.5)</td>
<td>0.25 (110.5)</td>
<td>0.22 (110.5)</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>$-8672.2$</td>
<td>$-8855.8$</td>
<td>$-11871.0$</td>
<td>$-8263.1$</td>
<td>$-10552.1$</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>$-121.0$</td>
<td>$-56.4$</td>
<td>$808.7$</td>
<td>$-272.2$</td>
<td>$468.2$</td>
</tr>
<tr>
<td>Observations</td>
<td>6110</td>
<td>6110</td>
<td>6110</td>
<td>6110</td>
<td>6110</td>
</tr>
</tbody>
</table>

Table 3.10: Statistics for the speed comparison for the deceleration case.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
<th>Model 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE ($m/s$)</td>
<td>1.8</td>
<td>1.9</td>
<td>1.0</td>
<td>1.9</td>
<td>1.2</td>
</tr>
<tr>
<td>MAPE (%)</td>
<td>123.2</td>
<td>126.8</td>
<td>29.9</td>
<td>137.2</td>
<td>39.4</td>
</tr>
<tr>
<td>ME ($m/s$)</td>
<td>1.23</td>
<td>1.21</td>
<td>$-0.04$</td>
<td>1.23</td>
<td>$-0.06$</td>
</tr>
<tr>
<td>U (fraction)</td>
<td>0.0050</td>
<td>0.0050</td>
<td>0.0031</td>
<td>0.0052</td>
<td>0.0037</td>
</tr>
</tbody>
</table>

3.5 Discussion

According to the results in the previous two sections, the developed model (Equation (3.6)) fits the acceleration case better than the deceleration one. In the former case, all model specifications are able to yield fairly desirable estimates while in the latter one, only the models which includes the parameter $p$ (Model 3 and Model 5) are to some extents able to capture the acceleration behavior. This can be caused by both the feature of the model function and the feature of the cyclists’ deceleration behavior. More work can be done in the future to address this particular issue in the deceleration case.
3.6 Summary

This chapter presents the investigation into the cycling performance and the modeling of cyclist behavior. A cycling regime is suggested at the beginning of the chapter as the cornerstone of the subsequent research. With the cycling behavior decomposed into three parts, including the acceleration, deceleration and cruising, corresponding sub-datasets are created. Descriptive analyses for all three parts are then conducted and illustrated in the chapter. Moreover, mathematical models were formulated to describe cyclists’ acceleration and deceleration behavior. The models were further estimated using the maximum likelihood estimator and evaluated by several goodness-of-fit measures.
Chapter 4

Conclusions and Future Work

4.1 Conclusions

In a relatively long term, the upward trend of cycling can be irreversible around the world given a number of facts, such as the increasing gasoline price, worsening traffic conditions and people’s growing pursuit of healthy lifestyle. This thesis, therefore, targets this prospective non-motorized transportation, presenting the developed methodologies for investigating cycling performance as well as for modeling cyclist behavior.

The thesis mainly comprises two parts of study, which are respectively outlined in chapter 2 and chapter 3. In general, chapter 2 demonstrates the whole efforts for data preparation, while chapter 3 details the investigation into the data and presents the developed models for cyclists’ acceleration and deceleration behaviors.

Chapter 2 starts with an introduction of the data collection. Naturalistic cycling data were gathered from August to October, 2013 in Stockholm. The GPS data documented by the Garmin EDGE 500 and Garmin Oregon were finally adopted for the research purpose. Different techniques were used to process and smooth cycling data profiles. For example, the locally weighted regression was applied to smooth the altitude profiles and the Kalman Filtering was used to cancel the measurement noise in GPS data. Furthermore, additional information, including the acceleration profiles and road gradient profiles, was derived from the processed speed profiles and processed altitude profiles, respectively.

In chapter 3, a cycling regime which divides the cyclist behavior into three parts, including the acceleration behavior, deceleration behavior and cruising behavior, is first suggested. Based on this regime, sub-datasets for each type of behavior were created and descriptive
analyses were also performed using the respective datasets prior to the model development. Some features of the cycling performance are thus figured out. For instance, during a cyclist’s acceleration process, the acceleration rate varies over time instead of remaining constant. More importantly, it starts and ends at zero and reaches the peak value in between, yielding an approximately U-shaped curve. On the other hand, the deceleration case turns out to be the mirror image of the acceleration case and shows quite similar features. Furthermore, key variables for such processes, e.g., the initial speed, final speed, average acceleration rate and maximum acceleration rate, are extracted and highlighted through plotted distributions. As for the cruising behavior, fundamental linear regressions were conducted to examine the relationship between the road gradient and cycling speed. Due to the limit of sample size, other social demographic variables are not taken into account in the current study.

The rest of chapter 3 is focused on the core of this thesis: the modeling of cyclists’ acceleration behavior. The ideas about modeling are mainly enlightened by the previous finding that cyclists’ acceleration profiles show approximately U-shaped curves. In this sense, a polynomial function which can yield U-shaped curves with an independent variable $x$ varying between 0 and 1 is adopted to formulate the general form of our acceleration model. A couple of model specifications are also presented in this chapter, with the model parameters estimated using the Maximum Likelihood Estimator (MLE). Besides, these specifications are evaluated and compared based on some goodness-of-fit measures, including RMSE, MAPE, ME, and U. Differences between the acceleration case and the deceleration case are also discussed.

4.2 Future Work

The current research can be mainly extended in three aspects. First, more efforts can be spent in the data processing part to make GPS data more accurate and reliable. This study highlights the tremendous ability of GPS data in terms of investigating travelers’ behavior from a microscopic perspective, but it also poses some issues in the application of GPS data. For example, in the current case, the cycling data was measured and documented with a time interval of one second, which is relatively too large for cycling activities since cyclists can averagely move around six meters per second according to our findings. As a result, information between each two seconds is missing and it may have a negligible influence on the final results. In order to solve this problem, techniques, such as map matching and
interpolation, can be implemented.

Second, much more data can be collected from cyclists with diverse social demographics. So far a series of methodologies have been developed to investigate and model cyclist behavior, yet variables such as age, physical conditions and so on, have not been examined whether they have significant effects on cyclist behavior. Particularly, the third model specification which incorporates person type variables (gender and bicycle type) might perform even better if some other omitted significant ones could be figured out.

Third, a bicycle traffic simulator can be developed to address practical issues, such as bicycle facility planning, bicycle sharing system optimization and so on. The suggested cycling regime can be used to manipulate cyclist behavior and the developed models in this study can be applied to simulate cyclists’ acceleration and deceleration processes.
Appendix A

Acceleration and Deceleration Profile Illustration
APPENDIX A. ACCELERATION AND DECELERATION PROFILE ILLUSTRATION

Acceleration Profile 1

Acceleration Profile 2

Speed Profile 1

Speed Profile 2

Acceleration Profile 3

Acceleration Profile 4

Speed Profile 3

Speed Profile 4
APPENDIX A. ACCELERATION AND DECELERATION PROFILE ILLUSTRATION

![Graphs of acceleration and speed profiles for different time intervals.](image-url)
Appendix B

MATLAB Code for Model Estimation and Validation

Note only the code for Acceleration Model 3 (which is also the general form of the developed model) is presented. The other models were estimated by using slightly modified codes based on this.
% Main function
clear all; clc;

% Data Selection
load ('0.5_acc.mat');
rand('state', 212); % a random seed

[EstimationData, ValidationData] = DataSelect(Acc_process); % a function which can randomly select 80% profiles for the model estimation and keep the rest for the validation

% Data specification
acc_observed = EstimationData(:,5);
n = size(acc_observed, 1);
X(:,1) = EstimationData(:,7); % Delta v
X(:,2) = EstimationData(:,4); % theta
X(:,3) = EstimationData(:,9); % Gender 1: Male 0: Female
X(:,4) = EstimationData(:,10); % Age_10−20 1: Yes, 0: No
X(:,5) = EstimationData(:,11); % Age_20−30 1: Yes, 0: No
X(:,6) = EstimationData(:,12); % Age_30−40 1: Yes, 0: No
X(:,7) = EstimationData(:,13); % Age_40−50 1: Yes, 0: No
X(:,8) = EstimationData(:,14); % Age_50−60 1: Yes, 0: No
X(:,9) = EstimationData(:,15); % Age_60−70 1: Yes, 0: No
X(:,10) = EstimationData(:,16); % RoadRacingBicycle 1: Yes, 0: No
X(:,11) = EstimationData(:,17); % MountainBicycle 1: Yes, 0: No
X(:,12) = EstimationData(:,18); % NormalCityBike 1: Yes, 0: No

% Estimation
% use mle to estimate sigma, define a new pdf with mu = 0 and sigma unknown
newpdf = @(x, sigma) (1/sqrt(2*pi*sigma*sigma)) * exp(-x.^2/2/sigma/sigma);

% Model 3
[P_model3, R_model3, J_model3, CovB_model3, MSE_model3, ErrorModelInfo_model3] = nlinfit(X, acc_observed, @accmodel_5, [0.5, 2, 1, 3]);

% P1: alpha, P2: beta, P3: p, P4: q
\[\text{DATA3} = \text{acc}_{\text{observed}} - P_{\text{model3}(1)} \cdot X(:,1) \cdot P_{\text{model3}(2)} \cdot X(:,2) \cdot P_{\text{model3}(3)} \cdot (1 - X(:,2) \cdot P_{\text{model3}(4)}).^2;\]

\[\hat{\sigma}_{\text{model3}} = \text{mle(DATA3, 'pdf', newpdf, 'start', 1);}\]

\[\text{acov}_{\text{model3}} = \text{mlecov(\hat{\sigma}_{\text{model3}}, \text{DATA3}, 'pdf', newpdf)};\]

\[t_{\sigma_{\text{model3}}} = \frac{\hat{\sigma}_{\text{model3}}}{\sqrt{\text{diag(\text{acov}_{\text{model3}})}}};\]

\[t_{P_{\text{model3}}} = \frac{P_{\text{model3}}}{\sqrt{\text{diag(CovB_{model3})}}};\]

\[\% t-value for the estimated sigma\]

\[\% t-value for the estimated parameters\]

\[\% \log\text{-likelihood}\]

\[\text{CorePart1} = \text{sum(\text{acc}_{\text{observed}}.^2)};\]

\[\text{CorePart2} = \text{sum((\text{acc}_{\text{observed}} - P_{\text{model3}(1)} \cdot X(:,1) \cdot P_{\text{model3}(2)} \cdot X(:,2) \cdot P_{\text{model3}(3)} \cdot (1 - X(:,2) \cdot P_{\text{model3}(4)}).^2).^2);}\]

\[\text{LL}_{0_{\text{model3}}} = -n/2 \cdot \log(2 \cdot \pi) - n \cdot \log(\hat{\sigma}_{\text{model3}}) - \frac{\text{CorePart1}}{2 \cdot \hat{\sigma}_{\text{model3}}^2};\]

\[\text{LL}_{B_{\text{model3}}} = -n/2 \cdot \log(2 \cdot \pi) - n \cdot \log(\hat{\sigma}_{\text{model3}}) - \frac{\text{CorePart2}}{2 \cdot \hat{\sigma}_{\text{model3}}^2};\]

\[\% \% \text{Model Validation}\]

\[\text{validate\_index} = \text{unique(ValidationData(:,6))};\]

\[\text{compare}_{\text{model3}} = [];\]

\[\text{temp\_size} = 0;\]

\[\text{for} \ i = 1: \text{size(\text{validate\_index},1)}\]

\[\text{temp\_position} = [];\]

\[\text{temp\_set} = [];\]

\[\text{temp\_position} = \text{find(ValidationData(:,6)==\text{validate\_index}(\text{i}))};\]

\[\text{temp\_set} = \text{ValidationData(\text{temp\_position},:)};\]

\[v_{\text{i}} = \text{temp\_set}(1,3);\]

\[v_{\text{f}} = \text{temp\_set(end,3);}\]

\[v_{\text{t}} = \text{temp\_set}(2,3);\]

\[\text{sp} = 1 + \text{temp\_size};\]

\[\text{ep} = \text{size(\text{temp\_set},1)} + \text{temp\_size};\]

\[\text{compare}_{\text{model3}}(:,5) = \text{ValidationData(:,6)};\]

\[\text{compare}_{\text{model3}}(\text{sp}:\text{ep},1) = \text{temp\_set(:,5)}; \% \text{real acceleration}\]
compare_model3(sp,ep,2) = temp_set(:,3); % real speed
compare_model3(sp,3) = 0; % column 3 predicted acceleration
compare_model3(sp,4) = v_i; % column 4 predicted speed
for j = 1:size(temp_set,1)
    theta = (v_t - v_i)/(v_f-v_i);
    acc_pre = P_model3(1) * temp_set(1,7).*P_model3(2).*
               theta.*P_model3(3).* (1 - theta.*P_model3(4)).^2;
    v_pre = v_t + acc_pre;
    compare_model3(sp+j,3) = acc_pre;
    compare_model3(sp+j,4) = v_t;
    v_t = v_pre;
end

end

end

compare_model3(end,:) = [];
[RMSPE_model3,RMSE_model3,MPE_model3,MAPE_model3,ME_model3],
U_model3] = ModelValidate(compare_model3(:,4),compare_model3(:,2),compare_model3(:,5))

%% Acc Data selection

function [EstimationData,ValidationData]=DataSelect(Acc_process)
cyclist_ID = unique(Acc_process(:,8));
% rng('default'); % make sure the results remain the same
y = 0;
for i = 1:size(cyclist_ID,1)
    position = [];
    acc_profile_id = [];
    select = [];
    position = find(Acc_process(:,8) == cyclist_ID(i));
    acc_profile_id = unique(Acc_process(position,6));
    t = round(0.8 * length(acc_profile_id));
    select = randperm(length(acc_profile_id),t);
    sp = 1 + y; % start point
ep = t + y; % end point
EstimationData_index(sp:ep) = acc_profile_id(select); % record the profile id
y = y + t;
end

% sort
EstimationData_index = sort(EstimationData_index);

% Dataset for estimation: 80% profiles
[LIA, LOCB] = ismember(Acc_process(:,6), EstimationData_index);
position_2 = find(LIA == 1);
EstimationData = Acc_process(position_2, :);

% Dataset for validation: 20% profiles
index = setdiff(Acc_process(:,1), EstimationData(:,1));
[LIA2, LOCB2] = ismember(Acc_process(:,1), index);
position_3 = find(LIA2 == 1);
ValidationData = Acc_process(position_3, :);
end

% Model Validation
function [RMSPE, RMSE, MPE, MAPE, ME, U] = ModelValidate(Y1, Y0, X)
N = size(Y1, 1);
RMSPE = sqrt(sum(((Y1 - Y0) ./ Y0).^2) / N);
RMSE = sqrt((1/N) * sum((Y1 - Y0).^2));
MPE = (1/N) * sum((Y1 - Y0) ./ Y0);
MAPE = (1/N) * sum(abs(Y1 - Y0) ./ Y0);
ME = (1/N) * sum(Y1 - Y0);
U = sqrt((1/N) * sum((Y1 - Y0).^2)) / (sqrt((1/N) * sum(Y1).^2) + sqrt((1/N) * sum(Y0).^2));
end
Bibliography


Modeling of Cyclists Acceleration Behavior Using Naturalistic Data

DING LUO