Analysis of Cyclist Behavior Using Naturalistic Data: Data Processing for Model Development

D. Luo[1] and X. Ma[1*]

[1] Intelligent Transport Systems Lab, Traffic and Logistics (ToL), KTH Royal Institute of Technology
Teknikringen 72, Stockholm 10044, Sweden

[*]Corresponding author e-mail: liang@kth.se

ABSTRACT

Cycling has been increasingly popular in many cities over the past decades because of its benefits for both environment and human health. However, there is still lack of knowledge on the characteristics specific to this traveler group and recent promotion of bicycle use in transport policies has even expanded the demand for understanding cyclist behavior and bicycle dynamics. It is believed that such understanding can further facilitate the evaluation and improvement of cycling safety as well as accessibility on the network. This paper therefore presents an essential methodological framework for processing and analyzing naturalistic data collected by commuter cyclists in Stockholm equipped with portable GPS devices. On one hand, the GPS coordinates are filtered by the Kalman smoothing algorithm to obtain accurate and consistent estimates of cyclists' position, speed and acceleration. On the other locally weighted regression is applied to abstract gradient profiles of cycling paths using data of both altitude and travel distance. After information estimation, the characteristics of cyclist acceleration behavior are then analyzed using statistical approaches. The results show that the acceleration profiles have a linear correlation with the total variance in speed during acceleration or deceleration. The data is finally applied to identify cyclist acceleration models proposed for the development of cycling simulation.

Keywords: cyclist behaviour, naturalistic cycling GPS data, information filtering, bicycle acceleration model and simulation.

1 INTRODUCTION

Many cities in Europe and US have witnessed the growing cyclist population over the past decades. Although the size of this community is still small in comparison to motorized vehicles, its high level of vulnerability has been recognized in traffic safety research. Meanwhile, new policy trends for sustainable transport development have also created a tremendous demand for more knowledge on the characteristics specific to this traveler group [1,2]. As the popularity of cycling keeps increasing in most urban areas, both traffic planners and policy makers are seeking useful analytical tools, which can assist in addressing bicycle-related planning and operational issues. The fact that the development of these tools, such as the bicycle traffic simulation models, highly depends on sufficient understanding of cyclist behavior makes it even more urgent for researchers to initiate related studies. To compensate for this shortage, this paper presents a study on the cyclist behavior based on naturalistic bicycling data. Inspired by the previous approach on driver behavior, the naturalistic data from commuter cyclists in

Stockholm are employed for the research. It is believed that the insight into the cyclist behavior illustrated in this study may promote the development of this economical, healthy and environmentally friendly mode of transport.

1.1 Literature Review

Naturalistic data has so far been widely accepted as the most accredited tool for analyzing road user behaviors. The trend of using this type of data started in the field of driver behavior studies around ten years ago [3-4], and now has been driving the development and evaluation of intelligent in-vehicle systems. Compared with traditional data sources, such as data from accident databases or data recorded at certain spots, naturalistic data is more capable of providing researchers with detailed and consistent information, thus contributing to better solutions to the problems.

Despite the prevalence of naturalistic data in driver behavior studies, a counterpart in bicycle-related studies has not been observed yet, with merely a few papers available to the public. For example, Johnson et al. [5] used videos filmed by cyclists' helmet-mounted cameras to identify risk factors for collisions or near-collisions involving on-road drivers and commuter cyclists. Parkin et al. [6] utilized naturalistic GPS data from ordinary cyclists to determine the design speed and acceleration for the cyclists in UK. Furthermore, Gustafsson et al. employed both GPS devices and video cameras [7]. Two data sources were combined in their developed data-analysis software, and issues as well as conflicts occurred during cyclists' trips were subsequently identified. More recently, newer and more intelligent elements have also been applied to the cycling data collection. For instance, by designing and using a well instrumented bicycle with multiple devices and sensors, Dozza et al. [8] gathered a large amount of naturalistic data which contained various information, such as the longitudinal acceleration, lateral acceleration and so on. In addition, a computer-based technique was developed, which can automatically perform cyclist data collection [9].

With various approaches of collecting cyclist data, however, deep analyses and models for cyclist behavior are still absent. Many studies only focus on cyclists' performance at intersections since it is known that vehicle-bicycle collisions occurred at intersections are the most common issues due to the insufficient clearance time for cyclists traveling at their ordinary cruising speed [10]. Pein [11] investigated cyclist performance both on multiuse trails and at three-leg intersections. The study found that the cyclists did not accelerate uniformly, with the acceleration rate decreasing after an initial increase. Nevertheless, the research was not continued and no more detailed results were available. Figliozzi et al. [12] developed a methodology for estimating cyclist acceleration and speed distributions at intersections. They employed a basic video setup to collect field data and further presented some statistical analyses on cyclist acceleration and cruising speed performance at intersections. In addition, some studies, e.g. [6], also tried to figure out whether cyclist demographics have an influence on cycling performance and cyclist behavior, yet no consensus so far has been reached.

1.2 Objectives and Structure

The main objective of this study is to evaluate the cycling characteristics of normal cyclists based on naturalistic data collected by portable GPS devices. The primary goals are to

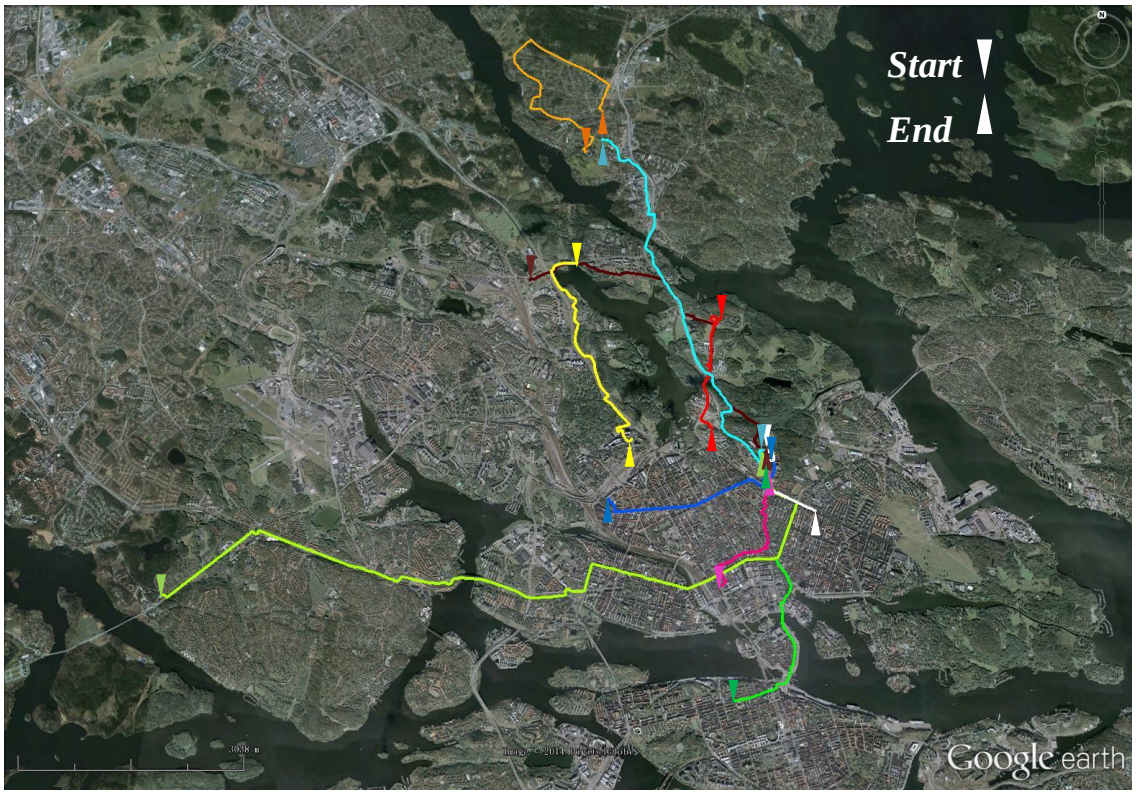
- Establish valid approaches to process raw data and estimate information of cyclist speed, acceleration and altitude;
- Analyze cyclist behavior (acceleration, deceleration and cruising) and infer its characteristics by statistical analysis;
- Show the potential of such data-driven approach in the application of modeling cyclist behavior.

The remainder of this article is organized as follow: in the second section the data collection and preprocessing are outlined, including the GPS data estimation and altitude data smoothing; then section 3 presents the methods and results in further data analysis; in the fourth sec-

tion, an application of naturalistic data in acceleration model is briefly introduced; finally, the research was summarized with conclusions drawn in the fifth section.

1. Garmin 60CSx **2. Garmin EDGE 500** **3. Garmin Oregon** **4. IPhone APP (Speed Tracker)**

(a)



(b)

Figure 1. A demonstration for four candidate devices (a); a demonstration for several representative cycling trajectories completed in the urban area of Stockholm, shown in Google Earth (b).

2 Data Collection and Preprocessing

2.1 Data Collection

The data collection so far has been organized twice in the autumn 2013 and the spring 2014 respectively. Eleven commuter cyclists in Stockholm, three female and eight male, were volunteered for the data collection. At the beginning four different types of devices were prepared for the data collection (see Figure 1a), yet practical testing results showed that the Garmin 60CSx and iPhone APP were neither capable of providing high quality data nor logging required information consistently. Therefore neither of them was employed for the data collection and all participants were later provided with handlebar mountable Garmin Edge 500 GPS devices except one who used his own Garmin Oregon GPS device. All participants were required to record their normal cycling trips as many as possible using the provided devices. Devices were regularly retrieved by the authors in order to upload the data. At last, 126 available cycling trips completed in the urban area of Stockholm were sorted out and the raw data was further stored and managed in a PostgreSQL database. Several representative cycling trajectories can be then viewed in Figure 1b.

Both applied Garmin devices were able to measure and record GPS data and altitude data with a time interval of one second depending on the high-sensitivity integrated GPS receivers and the internal barometric altimeters, respectively. The GPS data included not only the original measurements of latitude and longitude with eight digits in the fractional part, but also the derived information of distance and speed. Moreover, the altitude data turned out to be relative measurements instead of absolute measurements in most cases due to the limitation of the internal barometric altimeters.

2.2 GPS Data Estimation

The recorded GPS data contained information about the latitude, longitude, traveling distance and even instantaneous speed of each cyclist. It was, however, found that missing observations and noise involved turned out to be an obstacle for abstracting information for application. In fact, this issue is not rare in the tracking problems by GPS and other devices in which the loss of signal and noise inclusion have to be solved by signal processing technique.

During our data collection, participant cyclists were mostly traveling within the urban area, thus it is not unexpected that sometimes the functioning of GPS receivers were limited by the ambient buildings or shelters logging blank tuples in the data file. After trying some estimation approaches from the simple ARMA filters to more advanced adaptive filters, the Kalman filter (KF) based approaches finally brought the most consistent estimation results. The KF is an advanced method, which is able to give optimal estimates of discrete data [13]. The approach has been widely used in signal processing and control systems over the past few decades. In this study, application of the method allowed us to take advantages of high-frequency GPS measurements. As a result, not only the problem of missing observations could be addressed depending on the "Predict" procedure of the filter, but also the measurement noise in the GPS data could be largely removed simultaneously.

Kalman Filter based Method

The specific implementation of the KF in our case referred to a previous study [4], in which the extended Kalman smoothing algorithm, also called Rauch-Tung-Striebel smoother was applied to estimate the driver behavior data collected by an instrumented vehicle. In general, the state space model for the tracking problem can be written as follows:

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

$$\mathbf{Y}(t) = H \cdot \mathbf{X}(t) + \mathbf{W}(t) \quad (2)$$

where $\mathbf{X}(t)$ and $\mathbf{Y}(t)$, respectively, denote the state vector and measurement vector at time t; A denotes the state transition matrix; H denotes the relation matrix between the measurement and state vector; $\mathbf{V}(t)$ and $\mathbf{W}(t)$, respectively, denote the process noise and measurement noise and both of them are also assumed to be white noises.

Table 1 The Kalman smoothing algorithm

Forward Filtering

State update

$$\overline{\mathbf{X}}_{t|t-1} = A\widehat{\mathbf{X}}_{t-1|t-1}$$
$$\bar{P}_{t|t-1} = A\hat{P}_{t-1|t-1}A^T + Q$$

Measurement update

$$K_t = \bar{P}_{t|t-1}H^T(H\bar{P}_{t|t-1}H^T + R)^{-1}$$
$$\widehat{\mathbf{X}}_{t|t} = \overline{\mathbf{X}}_{t|t-1} + K_t(\mathbf{Y}_t - H\overline{\mathbf{X}}_{t|t-1})$$
$$\hat{P}_{t|t} = (I - K_tH)\bar{P}_{t|t-1}$$

Backward Smoothing

$$\widehat{\mathbf{X}}_{t|N} = \widehat{\mathbf{X}}_{t|t} + \Omega_t(\widehat{\mathbf{X}}_{t+1|N} - \widehat{\mathbf{X}}_{t+1|t})$$
$$\hat{P}_{t|N} = \hat{P}_{t|t} + \Omega_t(\hat{P}_{t+1|N} - \hat{P}_{t+1|t})\Omega_t^T$$
$$\Omega_t = \hat{P}_{t|t}A^T\hat{P}_{t+1|t}^{-1}$$

As Table 1 shows, the extended Kalman smoothing algorithm virtually consists of a conventional forward filtering process and a backward smoothing process. The backward process benefits from the offline smoothing in a sense that given the entire observation quence $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$, researchers can use the non-causal information to improve the estimation. For the forward filter, $\overline{\mathbf{X}}_{t|t-1}$ and $\bar{P}_{t|t-1}$, respectively, denote propagated state estimation (that is, the prediction of the state at t using observation before that instant) and its estimated covariance prior to time t; $\widehat{\mathbf{X}}_{t|t}$ and $\hat{P}_{t|t}$, respectively, denote the *a posteriori* state estimate and error covariance; K_t denotes the Kalman gain at time t. For the backward smoothing, $\widehat{\mathbf{X}}_{t|N}$ denotes the estimation of $\mathbf{X}_t$ given the whole data sequence $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_N$ is available; $P_{t|N}$ denotes the error covariance matrix.

Practical Results

Given the abundance in the available GPS information contained the latitude, longitude, distance and speed in the current case, the study decided to take advantage of it by performing two types of KF. Specifically, the distance and speed information was used for a one-dimensional filter while the latitude and longitude information for another two-dimensional filter. Both filters could then yield processed speed data as well as acceleration data, and further by comparing the results of two different filters the authors were also able to figure out whether the automatic derivation of distance and speed by the Garmin devices was accurate and which result from the filter was more suitable for subsequent research. This scenario turned out to be applicable in the current case owing to the high-accuracy latitude and longitude geodetic coordinates documented by the Garmin devices. There were eight digits in the fractional part for both latitude and longitude geodetic coordinates so that the authors could use them as the raw measurements for the KF.

The specific operation for the one-dimensional filter used the distance and speed information refers to the procedures outlined in [4] and the authors here will only illustrate details about the two-dimensional filter. One important thing that has to be pointed out before conducting the KF is that the geodetic coordinates (latitude and longitude) have to first converted to Cartesian coordinates, specifically Universal Transverse Mercator (UTM) [15] in the current case. This is because geodetic coordinates are not appropriate for the desired data processing so that only through this transformation can the position appear on a rectangular grid in the x-y format. The state space model was then formulated based on the physical state relation shown as follow:

$$s_x(t+1) = s_x(t) + v_x(t)\Delta t + \frac{1}{2}a_x(t)\Delta t^2 \quad (3)$$

$$s_y(t+1) = s_y(t) + v_y(t)\Delta t + \frac{1}{2}a_y(t)\Delta t^2 \quad (4)$$

$$v_x(t+1) = v_x(t) + a_x(t)\Delta t \quad (5)$$

$$v_y(t+1) = v_y(t) + a_y(t)\Delta t \quad (6)$$

Moreover, two independent random walk processes were created for the acceleration in order to complete the state space model. These random walk models worked in a way that the acceleration of next time stamp was determined by that of the current time stamp plus a random noise term denoted by $\varepsilon(t)$ which was assumed to be white noise. Models are shown as follows:

$$a_x(t+1) = a_x(t) + \varepsilon_x(t) \quad (7)$$

$$a_y(t+1) = a_y(t) + \varepsilon_y(t) \quad (8)$$

In summary, with Δt equal to 1s, the matrices in (1) and (2) can be written as follow:

$$\mathbf{X}(t) = \left[s_x(t)\ \ s_y(t)\ \ v_x(t)\ \ v_y(t)\ \ a_x(t)\ \ a_y(t)\right]^T;$$

$$\mathbf{Y}(t) = \left[\hat{s}_x(t)\ \ \hat{s}_y(t)\ \right]^T;$$

$$\mathbf{V}(t) = \left[\ 0\ \ 0\ \ 0\ \ 0\ \ \varepsilon_x(t)\ \varepsilon_y(t)\ \right]^T;$$

$$\mathbf{W}(t) = \left[\ \varepsilon_{s_x}\ \varepsilon_{s_y}\ \right]^T;$$

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix};$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & 1 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Assuming that all state variables were independent from each other, and that the covariance matrices were time-invariant, the covariance matrices Q and R became diagonal ones. Since the power levels of the noise term for the random walk models as well as geodetic coordinates were unknown, the filter was designed based on a principle that the performance of the filter is mainly determined by the ratio of the covariance matrices. By adjusting the ratio, more desirable smoothing results were obtained accordingly. The outcomes of both one-dimensional and two-dimensional filters are then compared in Figure 2. As Figure 2a shows, the trajectory resulted from the two-dimensional Kalman smoothing algorithm almost overlaps the measured one completely, indicating that this filter is highly reliable. Figure 2b and 2c, respectively, portray the smoothed speed profiles and derived acceleration profiles. Notably, the speed profiles resulted from two different filters (the red and green lines) are quite identical to each other. As for the acceleration profiles, the red curve representing the one-dimensional case in Figure 2c turns out to be smoother than the green one representing the two-dimensional case. Consequently, the result infers that one-dimensional filter is, though simple, already dependable enough. Hence the speed and acceleration profiles resulted from the one-dimensional Kalman filter was used for subsequent research.

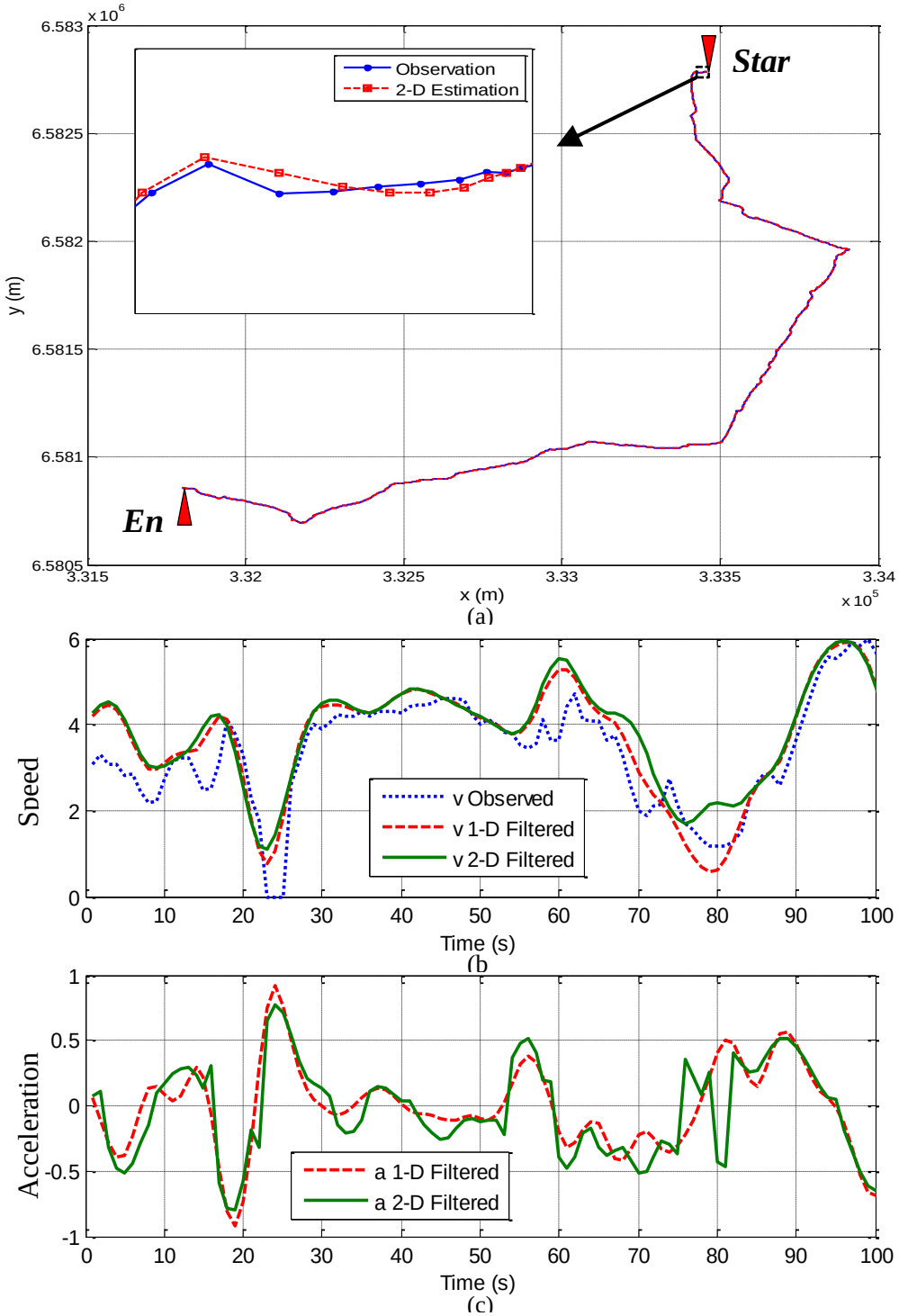


Figure 2. Comparison between the estimated and observed trajectories (a); Comparison between filtered speed profiles and observed profiles (b); Comparison between filtered speed profiles (c).

2.3 Altitude Data Smoothing

It has been mentioned that the present altitude data were relative observations in most cases. Moreover, its quality was not of high level either. For instance, measurement noises could be easily observed while plotting the original altitude profiles (see Figure 3b). Considering these drawbacks, it was therefore necessary to perform data smoothing on the altitude measurements so that the later derivation of gradient profiles could profit from the smoothed ones. At last, locally weighted regression [16] was applied to solve the smoothing problem in this study. Local regression is an approach that has so far been widely used to handle data filtering and smoothing. While performing local regression, a local curve to each point of interest is fitted using the observations around it. Extended from this very basic one, locally weighted regression then additionally takes into account a procedure of assigning weights for every involved neighbor point, so that important points can play more significant roles in the estimation.

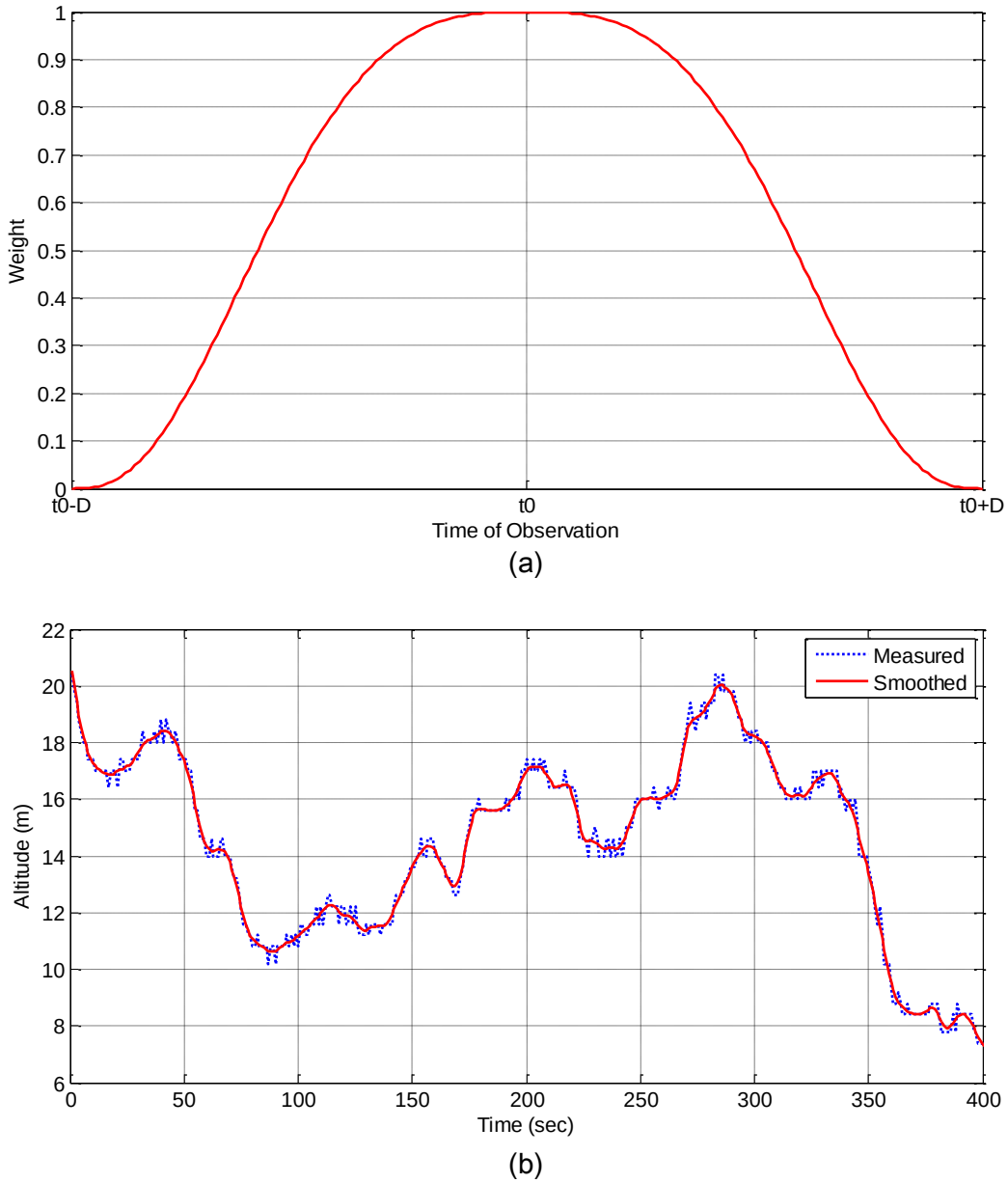


Figure 3. The tricube weight function (a); Comparison between the observed altitude data and smoothed altitude data (b).

Locally Weighted Regression

In the current application, the time series of altitude measurements for a single trip were denoted as: $x(t), t = 1,2, \dots, N$. The fitted value at $x(t)$ was then estimated as a polynomial fit to R observations (window) in the neighborhood of t_0 and the general form of polynomial function is specified by:

$$x(t) = f_{t_0}(t, \boldsymbol{\beta}_{t_0}) + \varepsilon_{t_0,t} \tag{9}$$

where $\boldsymbol{\beta}_{\mathbf{t_0}}$ denotes a vector of parameters to be estimated for the fitted curve; $\varepsilon_{t_0,t}$ denotes a normally distributed error term; $f_{t_0}(t, \boldsymbol{\beta}_{\mathbf{t_0}})$ denotes the fitted altitude at time t estimated by a local regression function centered at time t_0.

The weighted least squares estimation was then applied in order to obtain the parameters of local function $f_{t_0}(t, \boldsymbol{\beta}_{\mathbf{t_0}})$. Notably, the weights $w_{t_0}(t)$ assigned to t_0's neighbor points were computed based on the normalized time difference d (between t and the point of interest t_0). As d decreased, the weight of an observation in the window increased, implicating those points close to the target were more significant than those far away from the center. Furthermore, the following minimization problem was formulated. By solving this problem, a local altitude function centered at t_0 could be developed.

$$\min_{\boldsymbol{\beta}_{t_0}} [\mathbf{X}_{t_0} - f_{t_0}(t, \boldsymbol{\beta}_{t_0})]^T \mathbf{W}_{t_0} [\mathbf{X}_{t_0} - f_{t_0}(t, \boldsymbol{\beta}_{t_0})] \tag{10}$$

where $f_{t_0}(t, \boldsymbol{\beta}_{\mathbf{t_0}})$ denotes a corresponding vector of fitted values; $\mathbf{X}_{t_0}$ denotes a column vector of R observations adopted for the estimation; $\mathbf{W}_{t_0}$ denotes a $N \times N$ diagonal matrix of which elements coincide with the time difference based weights.

In the practical application of locally weighted regression, three components were required to be carefully considered and determined:

- The specification of the polynomial model $f_{t_0}(t, \boldsymbol{\beta}_{\mathbf{t_0}})$;
- Window size *R*;
- Specific weight assignment for all neighbor points.

The very basic linear polynomial model was finally adopted for the local regression since it could already provide good estimations for our study. The window size *R* was set to be 11, meaning 10 neighbor points (5 points each side) would be included in the local regression. As for the weight assignment, a tricube weight function recommended in literature [16] was used to calculate the weight for each neighbor point. This is because the weight function is not only smooth enough, but also meets the requirement that points closer to the center can be assigned higher weights. The weight function is specified below and illustrated in Figure 3a.

$$w(t_0, t) = (1 - d(t_0, t)^3)^3 \tag{11}$$

$$d(t_0, t) = \frac{|t_0 - t|}{D} \tag{12}$$

where $w(t_0, t)$ denotes the weight assigned to the observation $x(t)$ in the neighborhood of t_0; $d(t_0, t)$ denotes the time difference between t_0 and t; D denotes the number of points incorporated in the regression each side.

Figure 3b shows an example of smoothed altitude curve in comparison with the original altitude measurements. According to the plot, the measurement noise is appropriately removed while the profile becomes much smoother than before.

Gradient Profile Derivation

Considering the identified influence of road gradient on cycling performance as well as cyclist behavior, the study decided to derive gradient profiles using some available measurements including the distance and altitude. Specifically, the gradient was computed by dividing the change in altitude by the change in distance. Moreover, a criterion for the minimum moving distance (10 meters) was used as well, which implies that only when a part of trip covers

longer than 10 meters would the road gradients be calculated for the time stamps involved. In addition, the initial gradient g_1 was set to be 0 by default.

3. DATA ANALYSIS

3.1 Profile Selection

In the present study, an assumption that the cyclist behavior mainly consists of acceleration, deceleration and cruising behaviors was then employed. It could be then inferred based on this assumption that during a cycling process the cyclist always endeavors to reach and maintain his or her desired speed also varying on multiple factors, not only external ones (the road gradient and so on) but also internal ones (the cyclist's age, gender and so on). It is worth mentioning that such assumption refers to some driver behavior studies [17] but now similar concepts can also be seen in cyclist behavior studies [11-12] dealing with the estimation of cyclist acceleration and speed at intersections.

The main objective for data classification in the present study was to realize the identification of acceleration and deceleration profiles. The first step was to distinguish consecutively speeding-up and slowing-down data clusters and these identified clusters would then be regarded as the candidate profiles. Subsequently more strict criteria were considered to further eliminate insignificant profiles. These criteria include:

- The maximum acceleration of a profile a_{max} is smaller than $2\ \mathrm{m/s^2}$ (or the minimum acceleration a_{min} is greater than $-2\ \mathrm{m/s^2}$ for the deceleration case);
- The final speed of a profile v_f is smaller than $15\ \mathrm{m/s}$ (or the initial speed v_i is lower than $15\ \mathrm{m/s^2}$ for the deceleration case);
- The acceleration time of a profile t_a is between 4s and 15s;
- The total distance during the acceleration process d_a is not shorter than 5 meters;
- The road gradient over the entire acceleration process g_a is neither greater than 10% nor smaller than -10%;
- The change in speed of an acceleration process is significant. Specifically, an index τ defined as follow is not lower than 50%.

$$\tau = \frac{|v_f - v_i|}{\max(v_f, v_i)} \times 100\% \tag{13}$$

As a result, a total of 839 acceleration profiles and 613 deceleration profiles were respectively selected for analysis.

3.2 Descriptive Analysis

While examining the acceleration and deceleration profiles, it was found that the speed-time profiles for both acceleration and deceleration cases demonstrated S-shape curves (see Figure 4c and 4d) and the S-shape speed-time curves further resulted in U-shape acceleration-time curves as illustrated in Figure 4a and 4b. Based on these characteristics, a couple of critical variables were then extracted or computed for every single profile. These variables include:

- Initial speed v_i;
- Final speed v_f;
- Incremental speed Δv (or declined speed for the deceleration case);
- Average acceleration rate a_{avg} (or average deceleration rate d_{avg} for the deceleration case);
- Maximum acceleration rate a_{max} (or maximum deceleration rate d_{max} for the deceleration case);
- Acceleration time t_a.

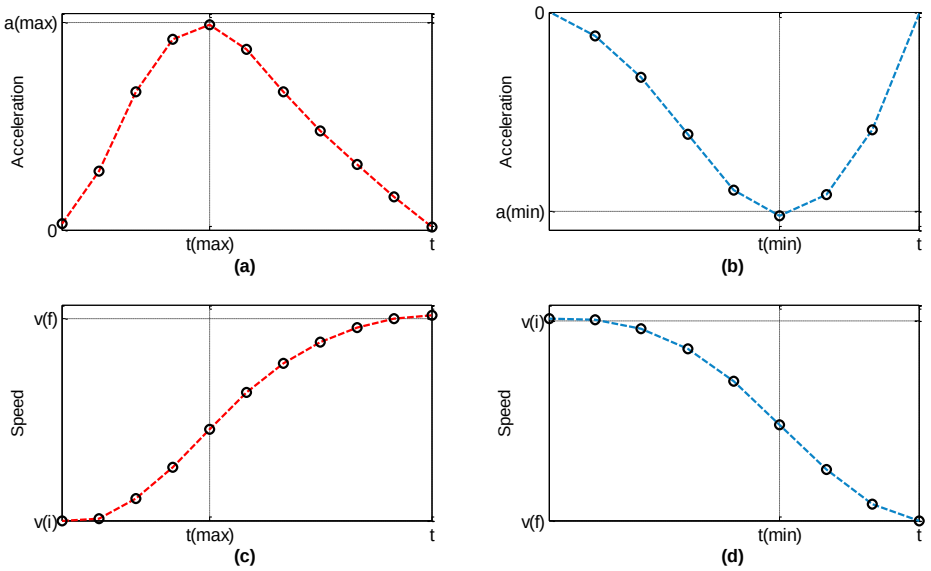


Figure 4. Typical acceleration-time and speed-time profiles for both acceleration ((a) and (c)) and deceleration cases ((b) and (d)). $\boldsymbol{v(i)}$ and $\boldsymbol{v(f)}$, respectively, denote the initial speed and final speed of a process.

Figure shows the distributions of these variables for both acceleration and deceleration cases. It can be seen both cases resemble to each other except for the sequence (initial speed and final speed) and the sign (acceleration rate). These distributions intuitively reveal some detailed characteristics of cyclist behavior. For example, cyclists hardly accelerated at a relatively high speed (greater than 4 m/s) and in most cases their speed increased by around 4 m/s. Most maximum acceleration rates were lower than 1 m/s^2 while the average maximum acceleration rate was even smaller (mostly lower than 0.5 m/s^2). Moreover, the acceleration time in most cases is shorter than 10 seconds.

3.3 Statistical Analysis

Correlation and regression analysis were further applied to all variables and it started with the calculation of correlation coefficient, which accounts for the power level of the linear relationship between two variables. The equation is shown as follow:

$$\rho_{X,Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{14}$$

where μ_X denotes the mean of variable *X*; σ_X denotes the standard deviation of variable *X* (the same for variable *Y*).

In the practical application, instead of the population correlation coefficient, the sample one was adopted and it is calculated as follow:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_{\square}}\right)\left(\frac{Y_i-\bar{Y}}{s_y}\right) \tag{15}$$

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

According to the analysis, the variable Δv representing the variation of speed in the acceleration or deceleration case had strong linear correlations to both maximum and average acceleration rates. In the acceleration case, the correlation coefficients between the incremental speed and the maximum acceleration rate and average acceleration rate respectively reached 0.7882 and 0.7664, whereas these two values were 0.7806 and 0.7628 in the deceleration case. Besides, the analysis did not find any significant linear correlation between the road

gradient and other variables involved in the acceleration process. All related correlation coefficients turned out to be quite small (around 0.1).

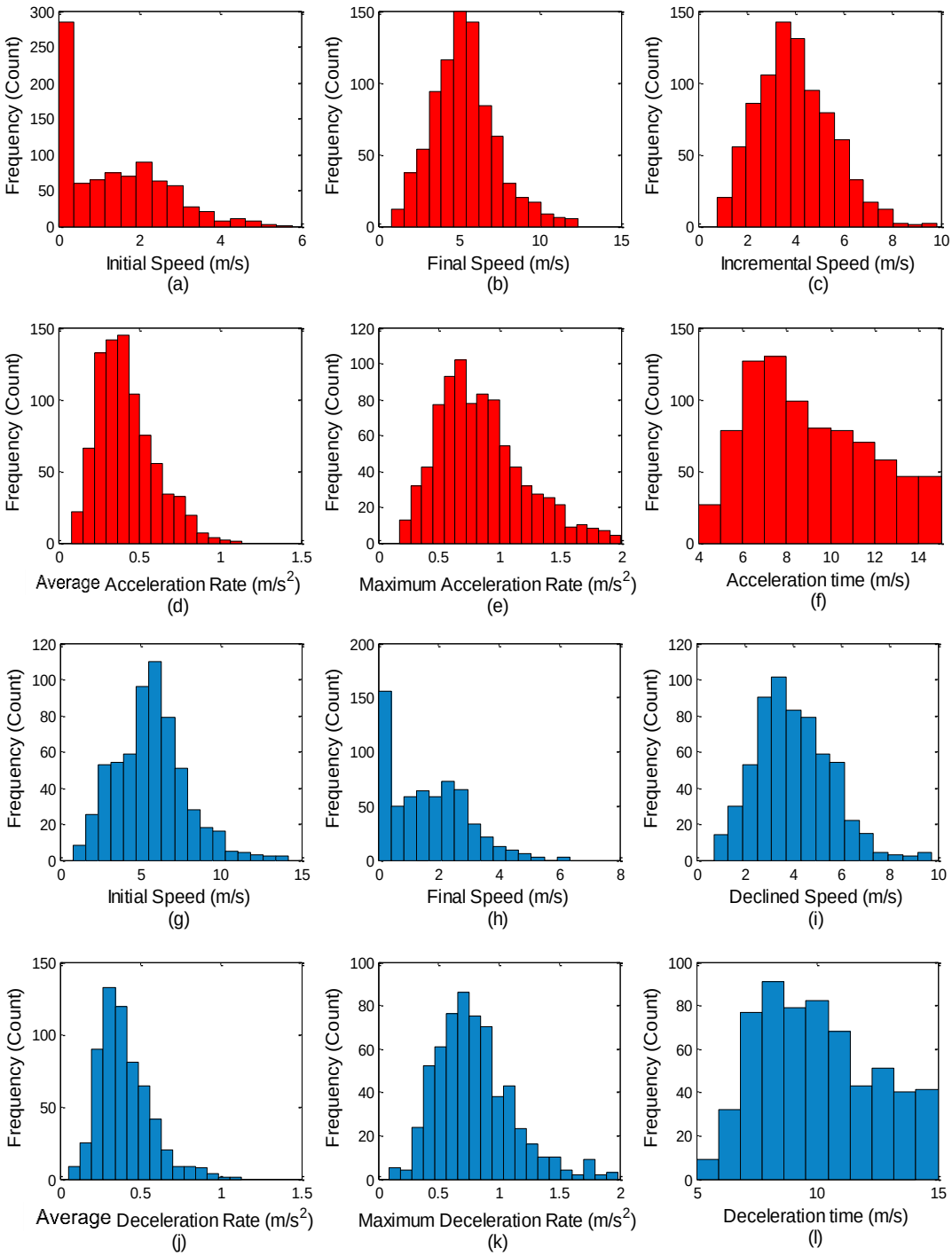


Figure 5. Distributions of significant variables for cyclists' acceleration and deceleration performance**.**

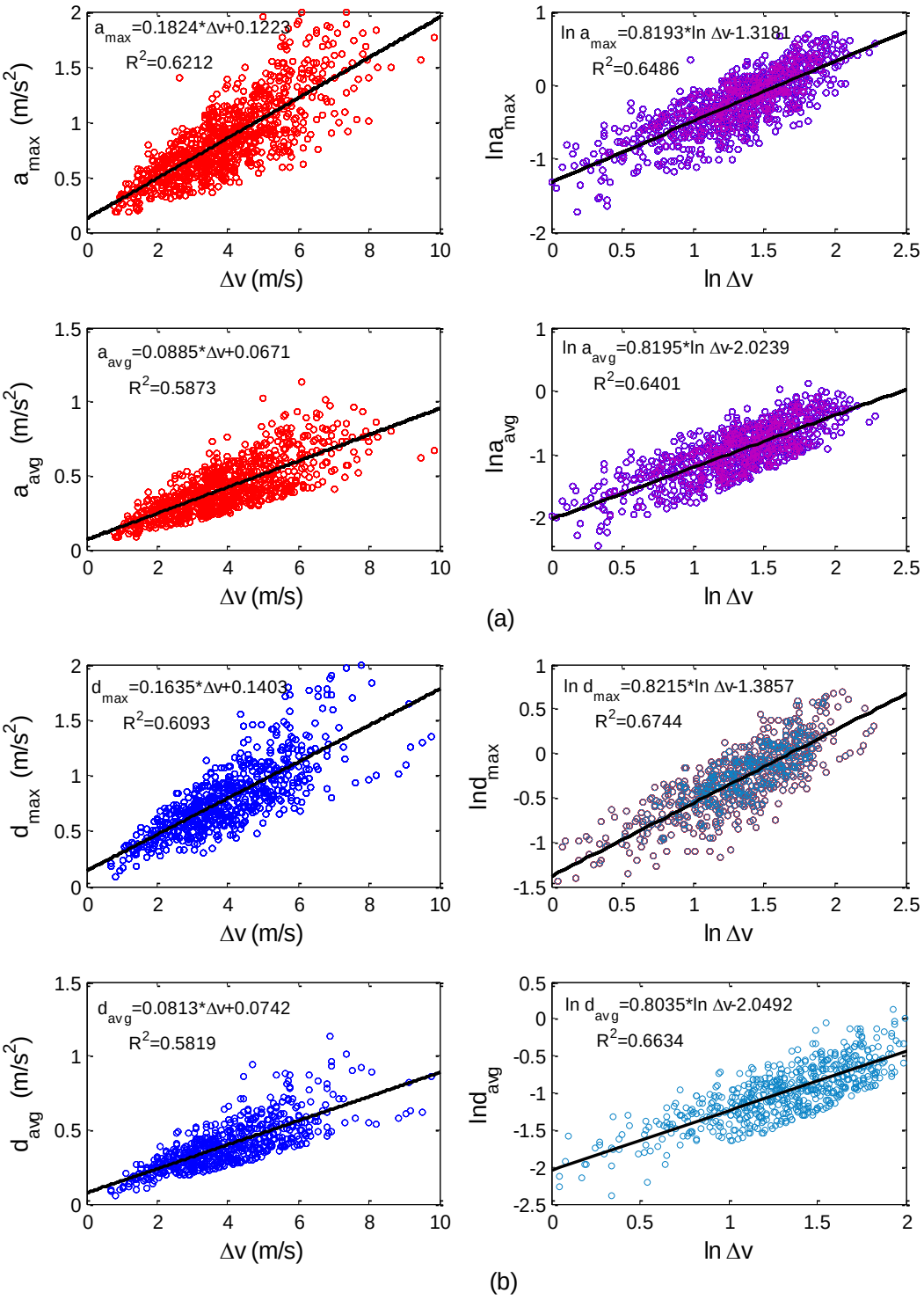


Figure 6. Comparison between the regression model with the original data and the model with the transformed data (Natural logarithm); acceleration case (a); deceleration case (b).

Figure 6 shows the scatter plots involving three main variables (Δv, a_{avg} and a_{max}). Notably, quite similar patterns can be observed in both acceleration and deceleration cases. For example, as Δv increases in the acceleration case, the maximum acceleration rate and average acceleration rate correspondingly increase. However, the oscillation of both acceleration variables can be clearly observed in the meanwhile, especially when the variation in Δv is significant, say, between six and ten. This particular observation hence implies the existence of

heteroscedasticity in these linear relationships. The corrections for heteroscedasticity were eventually achieved by transforming the variables to their natural logarithm. Comparisons can be seen in Figure 6.

4. APPLICATION IN ACCELERATION MODELLING

The naturalistic patterns collected by different methods have been applied to evaluate cycle traffic performance with respects to safety, mobility and accessibility [5,7,8]. In the meantime, similar types of data collected for motorized vehicles become a popular approach for behavior modeling in advanced traffic analysis. Nevertheless, unlike the importance of naturalistic data for basic driver behavior study, application of such data in cyclist behavior modeling hasn't been found in literature. One of the objectives of present research therefore is to overcome this limitation and demonstrate the applicability of such data.

Using the processed cyclist GPS information, the project further developed a mathematical approach to model the acceleration/deceleration behavior of this traveler group. The basic idea follows the aforementioned cycling regime which assumes that the cyclist behavior is comprised of acceleration, deceleration and cruising, and a cyclist will keep accelerating or decelerating before he or she reaches his or her desired cruising speed. Since the data analysis has figured out that both acceleration and deceleration profiles have U-shape curves despite the wide variation in individual and external factors (see Figure 4), the cyclist acceleration behavior can be modeled depending on the feature. Such way of modeling is also seen in driver behavior studies. For example, Akçelik et al. [19] proposed an acceleration modeling approach taking advantage of the characteristics of drivers' acceleration profiles.

In our model development process, to capture the U-shape of acceleration process became one of the most essential points. The model proposed in our research applied instantaneous speed $v(t)$ as the independent variable and other factors playing roles in adjusting estimation results were also employed for the mathematical formulation. In general, a nonlinear acceleration model was developed as follow:

$$a(t) = \alpha \cdot \Delta v^{\beta} \cdot \theta(t)^{n}(1 - \theta(t)^{m})^{2} + \varepsilon(t) \tag{16}$$

$$\theta(t) = \frac{v(t) - v_i}{v_f - v_i} \tag{17}$$

where *a(t)* denotes the acceleration rate at time *t* being modeled; α, n, m and β are essential model parameters to be estimated; Δt denotes the incremental (or declined) speed; $\varepsilon(t)$ denotes the i.i.d random term which follows a statistical distribution (e.g. normal) with zero mean. Given the specific assumption of this random term, the model was subsequently estimated using the Maximum Likelihood Estimation (MLE) approach. In fact, the proposed model follows the stimulus-response principle similar to the GM car-following models [20]. Specifically, the requirement of adapting the initial speed to the cyclist's desired speed decides how much he or she needs to accelerate, while the polynomial of $\theta(t)$ makes the acceleration process fulfill a U-shape curve.

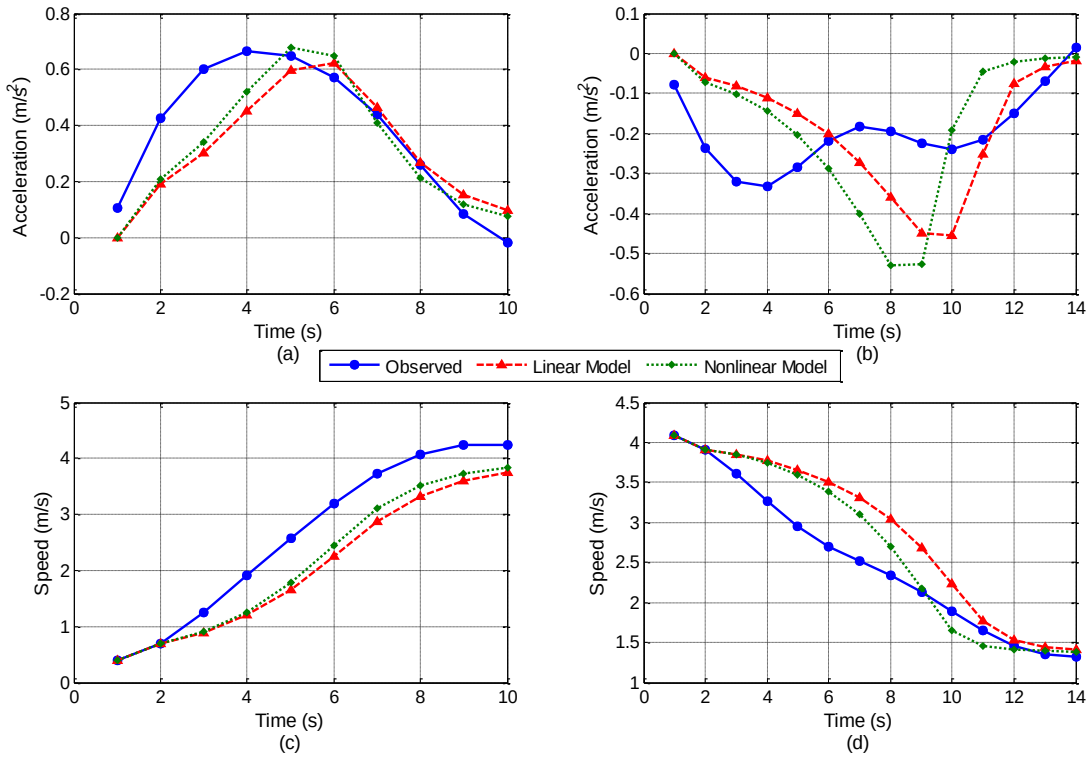


Figure 7. Comparison between the observed profiles and estimated acceleration-time profiles and speed-time profiles for both acceleration ((a) and (c)) and deceleration cases ((b) and (d)).

Previously selected acceleration and deceleration profiles were randomly divided and used for the model estimation (80%) as well as validation (20%). Examples of close-loop simulation results for two basic models with β fixed to be 1 (Linear) or not (Nonlinear) are then demonstrated in Figure 7.

5. SUMMARY AND CONCLUSIONS

A good insight into cyclist behavior has never been so desired as now since increasing endeavor has been spent to improve the convenience as well as safety of this economical, healthy and environmentally friendly mode of transport. Given the trend of analyzing the behavior of road users depending on naturalistic data, researchers have built capacity to obtain more detailed and accurate characteristic data on bicycle dynamics and cyclist behavior.

This paper presents specific methodologies for collecting, processing and analyzing naturalistic data from commuter cyclists in Stockholm who are provided with portable GPS devices. Moreover, an application of processed naturalistic data is also briefly illustrated at last to reflect the high level of its usability. In the present study, the naturalistic data comes from two sources: a high-sensitivity integrated GPS receiver and an internal barometric altimeter. Collected raw data is processed by applying the Kalman smoothing algorithm to more reliable GPS data, and locally weighted regression to more consistent altitude data, respectively. Information in cycling speed and acceleration rate is in the meanwhile estimated, and gradient profiles are derived by using both altitude and distance data.

Furthermore, given the assumption that the cyclist behavior is comprised of acceleration, deceleration and cruising, dedicated profiles for acceleration and deceleration are identified based on the criteria empirically established. Through examining all these profiles, the authors find that cyclists do not accelerate uniformly. Instead, the acceleration-time profiles show U-shape curves with both the initial acceleration rate and final acceleration rate being zero. On the basis of this property, some interesting characteristics of the acceleration behavior are revealed by using statistical approaches. One important result shows that the cyclists' acceleration has a linear correlation with the variance in speed over the acceleration process. This find-

ing then promotes the subsequent analytical work in which a stimulus-response type of acceleration model is proposed to describe cyclist behavior. While the modeling of cyclists' acceleration behavior turns out to be the major application of the current research, this paper demonstrates the potential of the collected naturalistic cycling data for identification of behavioral models. It is expected in our future study that more factors will be considered for behavioral modeling and therefore more naturalistic data will have to be collected.

ACKNOWLEDGMENTS

The authors would like to acknowledge the Swedish County Insurance foundation (Länsförsäkringar, Project P13/12) and J. Gust. Richert foundation for funding this research.

REFERENCES

[1] European Commission, *Traffic Safety Basic Facts 2012*. DaCoTA-Project, 2013.

[2] Karsch, H. M., Hedlund, J. H., Tison, J., and Leaf, W. A. *Review of Studies on Pedestrian and Bicyclist Safety, 1991-2007*, No. DOT HS 811 614, 2012.

[3] Dingus, T. A., S. G. Klauer, V. L. Neale, A. Petersen, S. E. Lee, and J. Sudweeks. *The 100-Car Naturalistic Driving Study. Phase II: Results of the 100-Car Field Experiment*, HS 810 593. Virginia Tech Transportation Institute, Virginia Polytechnic Institute and State University, Blacksburg, Va.; NHTSA, U.S. Department of Transportation, Washington, D.C., 2006.

[4] Ma, X., and Andréasson, I. Behavior Measurement, Analysis, and Regime Classification in Car-Following. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 8, No. 1, 2007, pp. 144-156.

[5] Johnson, M., Charlton, J., Oxley, J., and Newstead, S. Naturalistic Cycling Study: Identifying Risk Factors for On-road Commuter Cyclists. *Annals of Advances in Automotive Medicine*, Vol. 54, 2010, pp. 275-283.

[6] Parkin, J., and Rotheram, J. Design Speeds and Acceleration Characteristics of Bicycle Traffic for Use in Planning, Design and Appraisal. *Transport Policy*, Vol. 17, No. 5, 2010, pp. 335-341.

[7] Gustafsson, L., and Archer, J. A Naturalistic Study of Commuter Cyclists in the Greater Stockholm Area. *Accident Analysis & Prevention*, Vol. 58, 2013, pp. 286-298.

[8] Dozza, M., and Fernandez, A. Understanding Bicycle Dynamics and Cyclist Behavior from Naturalistic Field Data. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 15, No. 1, 2014, pp. 376-384.

[9] Zaki, M. H., Sayed, T., and Cheung, A. Computer Vision Techniques for the Automated Collection of Cyclist Data. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2387, Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 10-19.

[10] Wachtel, A., Forester, J., and Pelz, D. Signal Clearance Timing for Bicyclists. *ITE journal*, Vol. 65, No. 3, 1995, pp. 38-38.

[11] Pein, W. Bicyclist Performance on a Multiuse Trail. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 1578, Transportation Research Board of the National Academies, Washington, D.C., 1997, pp. 127-131.

[12] Figliozzi, M., Wheeler, N., and Monsere, C. M. Methodology for Estimating Bicyclist Acceleration and Speed Distributions at Intersections. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2387, Transportation Research Board of the National Academies, Washington, D.C., 2013, pp. 66-75.

[13] Kalman, R. E. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, Vol. 82, No. 1, 1960, pp. 35-45.

[14] Haykin, S. S. (Ed.). *Kalman Filtering and Neural Networks*. Wiley, Inc., New York, 2001, pp. 123-174.

[15] Hager, J. W., Behensky, J. F., and Drew, B. W. *The Universal Grids: Universal Transverse Mercator (UTM) and Universal Polar Stereographic (UPS)*, Edition 1. Topographic Center, Washington D.C., 1989.

[16] Cleveland, W. S., and Devlin, S. J. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, Vol. 83, No. 403, 1988, pp. 596-610.

[17] Kosonen, I. *HUTSIM: Urban Traffic Simulation and Control Model: Principles and Applications*. Doctoral Thesis, Helsinki University of Technology, 1999, pp. 62.

[18] Luo, D. *Modeling of Cyclists Acceleration Behavior Using Naturalistic Data*. M.Sc. Thesis. KTH Royal Institute of Technology, 2014.
[19] Akçelik, R., and Biggs, D. C. Acceleration Profile Models for Vehicles in Road Traffic. *Transportation Science*, Vol. 21, No. 1, 1987, pp. 36-54.
[20] Gazis, D. C., Herman, R., and Potts, R. B. Car-following Theory of Steady-State Traffic Flow. *Operations Research*, Vol. 7, No. 4, 1959, pp. 499-505.