Information **2015**, *6*, 212-227; doi:10.3390/info6020212

OPEN ACCESS

information

ISSN 2078-2489

www.mdpi.com/journal/information

Article

Identifying Travel Mode with GPS Data Using Support Vector Machines and Genetic Algorithm

Fang Zong *, Yu Bai, Xiao Wang, Yixin Yuan and Yanan He

College of Transportation, Jilin University, Changchun 130022, China;
E-Mails: 274048322@qq.com (Y.B.); 1411107559@qq.com (X.W.); 370828980@qq.com (Y.Y.); heyn14@mails.jlu.edu.cn (Y.H.)

* Author to whom correspondence should be addressed; E-Mail: zongfang@jlu.edu.cn; Tel./Fax: +86-431-8509-4779.

Academic Editors: Baozhen Yao and Yudong Zhang

Received: 15 April 2015 / Accepted: 27 May 2015 / Published: 4 June 2015

Abstract: Travel mode identification is one of the essential steps in travel information detection with Global Positioning System (GPS) survey data. This paper presents a Support Vector Classification (SVC) model for travel mode identification with GPS data. Genetic algorithm (GA) is employed for optimizing the parameters in the model. The travel modes of walking, bicycle, subway, bus, and car are recognized in this model. The results indicate that the developed model shows a high level of accuracy for mode identification. The estimation results also present GA's contribution to the optimization of the model. The findings can be used to identify travel mode based on GPS survey data, which will significantly enhance the efficiency and accuracy of travel survey and data processing. By providing crucial trip information, the results also contribute to the modeling and analyzing of travel behavior and are readily applicable to a wide range of transportation practices.

Keywords: Global Positioning System (GPS); travel survey; travel mode; Support Vector Classification; genetic algorithm

1. Introduction

Travel surveys are one of the most important ways of obtaining critical information needed for transportation planning and decision making. Traditionally, a travel survey used to be conducted using different methods such as telephone/face-to-face interviews and computer-based reporting to maintain a

diary [1]. These have proven to be a burden for participants to use, as well as being expensive and time consuming [2]. In addition, in these surveys, respondents often miss short trips, and round up the travel mode and travel time. The GPS-based travel survey can address this problem by recording travelers' GPS traces automatically. With the advantages of reducing respondents' participant burden and increasing data accuracy, GPS-based travel surveys have been conducted in many large cities, such as Beijing and New York. Nevertheless, GPS records cannot provide us with trip information that can be applied directly when analyzing and modeling travel behavior. In order to obtain the needed trip information, several major modeling steps, e.g., trip identification, mode detection, and travel time determination, have to be conducted based on the raw GPS data. This paper will present a model for travel mode detection, which is one of the crucial steps in trip information identification. Compared with the previous studies, it will try to enhance the identification accuracy by employing the SVC as well as using more GPS records.

The remainder of this paper is organized as follows. In Section 2, a review of identification of travel mode with GPS data in general is presented. Section 3 is a description of the data available for the study. This section is followed by a presentation of SVC and GA in Section 4. Section 5 presents the constructing process and modeling results of the mode detection model. The paper closes with some major conclusions and a discussion of future research directions.

2. Existing Literatures

Being a crucial step in GPS-based travel behaviors identifying, mode recognition has been investigated by many studies. Some researchers designed a threshold for parameters to detect travel mode, which is also called the criteria-based method. For example, Gong *et al.* [3] detected walking, subway, bus, and car by setting criteria for variables of speed and acceleration, *etc*., with the GPS data collected in New York. Liu and Zheng [4] used parameters including average speed and median speed to identify walking, bicycle, car, bus, and subway. Besides GPS data, some studies also introduced Geographic Information System (GIS) technology to recognize travel modes, especially bus and subway. For example, GIS and actual road network information were utilized in Stopher and Greaves's study [1] for bus and subway recognition. Bohte and Maat [5] introduced a method combining GPS data, GIS information, and a web-based validation application to identify walking, bicycle, car, and railway.

Most of the previous studies conducted mode identification by building detection models. A lot of methods in this category, including neural network [6,7], decision tree [8–10], Bayesian network [9,11,12], Support Vector Machines (SVM) [9,13,14], and conditional random field [9], have been applied in detecting travel mode. Most of the input variables come from the GPS data itself. In these studies, Zhang *et al*. [13] presented an SVM model, which achieved a relatively higher average accuracy rate, which proves the better performance of SVM than other methodologies in mode detection.

Although the previous studies proposed many methods for mode identification, most of these studies' sample sizes were quite small. For example, Zhang [15] used the GPS data collected from 23 volunteers, Gong *et al.* [3] employed 35 respondents' GPS traces, Vij *et al.* [16] conducted the mode detection by using 11 travelers' GPS data, and Du and Aultman-Hall [17] monitored 12 volunteers' GPS traces. The inadequate data sample size may limit the procedure of model construction and estimation, and turn out to impact the performance of mode identification. Moreover, mode identification in big cities is more

complicated due to complex traffic conditions and infrastructures, like urban canyons, bus routes, and subway networks. A large sample size can benefit data processing so as to provide more records for data analysis and model constructing. Bolbol *et al.* [18] took advantage of the GPS data of a large-scale study conducted in the Netherlands in 2007, which collected 1104 respondents' one-week survey data. Their results indicated that, for mode detection in a one-day GPS-based travel survey, the sample size needed is 289 and 271, respectively, for bus and car trips. By enhancing the sample size of GPS data, the accuracy rate of mode detection can be further enhanced. Therefore, this paper will construct the mode detection by using GPS survey data collected from 900 respondents, which is also part of a large-scale OD survey in Beijing, 2010. In order to enhance the performance of mode identification, SVC will be employed to establish the detection model.

3. Data

In this section the major information on survey methods and data processing is introduced. Beijing is chosen as the study area of this paper. Being the capital city of China, it has an integrated urban land use and composite transportation network, which includes a complex road network and one of the largest public transit systems in the world. The composite land use and transportation conditions make the data survey and mode detection very complicated. The problem of signal loss due to urban canyon and subway trips also has to be addressed in data processing.

3.1. Data Survey

This study takes advantage of GPS survey data collected as part of a large-scale OD survey conducted in Beijing in 2010. The data sample consists of 900 respondents' more than one million travel records collected during the survey period from 20 October 2010 to 26 November 2010. Each respondent has an average of about 1100 records. The survey area was 16,410.54 km^2 and the residing population of Beijing in 2010 was about 19.6 million. During the survey period, in addition to one-day GPS records, the 900 respondents also reported their travel diaries and socio-demographic information by filling in a paper form. With records containing missing values eliminated, our final sample consists of 1,872,431 GPS records.

Each record in the dataset represents a GPS signal that was captured consecutively at the 5-s interval by the GPS device (i-gotU GT-600) and contains information on index, date and time (universal coordinated time, UTC), latitude, longitude, altitude (m), speed (m/h), course (°), distance (m), EHPE (estimated horizontal position error, cm), and satellite ID.

3.2. Data Processing

For each GPS record, the first step of data processing is to change the geodetic coordinates (latitude and longitude) into geographical coordinates (X and Y coordinate) as well as the time and date in UTC value into local time. Being crucial measures for examining the quality of GPS records, satellite ID, EHPE, speed, and position (altitude, longitude, and latitude) are applied to remove invalid records in the dataset. Then GPS points are combined into trips and activities. Zong *et al.* [19] show the detailed process of identifying trips and activities with GPS data. Consisting of four sub-steps, namely, dividing

status segments, identifying activities, recognizing trips, and determining intermediate stops, the identification process can determine trips and activities based on GPS logs with a high level of accuracy [19].

Eight typical identification parameters, as shown in Table 1, concerning the speed, acceleration, travel time, and trip distance of each trip, are then employed to represent the mode characteristics based on previous studies [7]. Then, the threshold ranges for each identification parameter are defined, according to the traffic condition and complex building environment in Beijing as well as the results that previous studies presented [20]. The thresholds of identification parameters (shown in Table 1) are then used to filter the data samples. After data filtering, the final sample consists of 85,120 GPS trips.

Table 1. The thresholds of identification parameters.

Identification Parameters	Threshold Ranges
Average speed (km/h)	1–110
Maximum speed (km/h)	1–110
75th percentile of speed (km/h)	0–110
Acceleration (km/h^2)	0–36,000
75th percentile of acceleration (km/h^2)	0–36,000
Travel time (h)	0–10
Trip distance (km)	0–1000
Standard deviation of speed	0–100

The daily activity and travel pattern of respondent #010 are shown in Figure 1. The major trip characteristics as well as the starting and ending time of each trip are also presented.

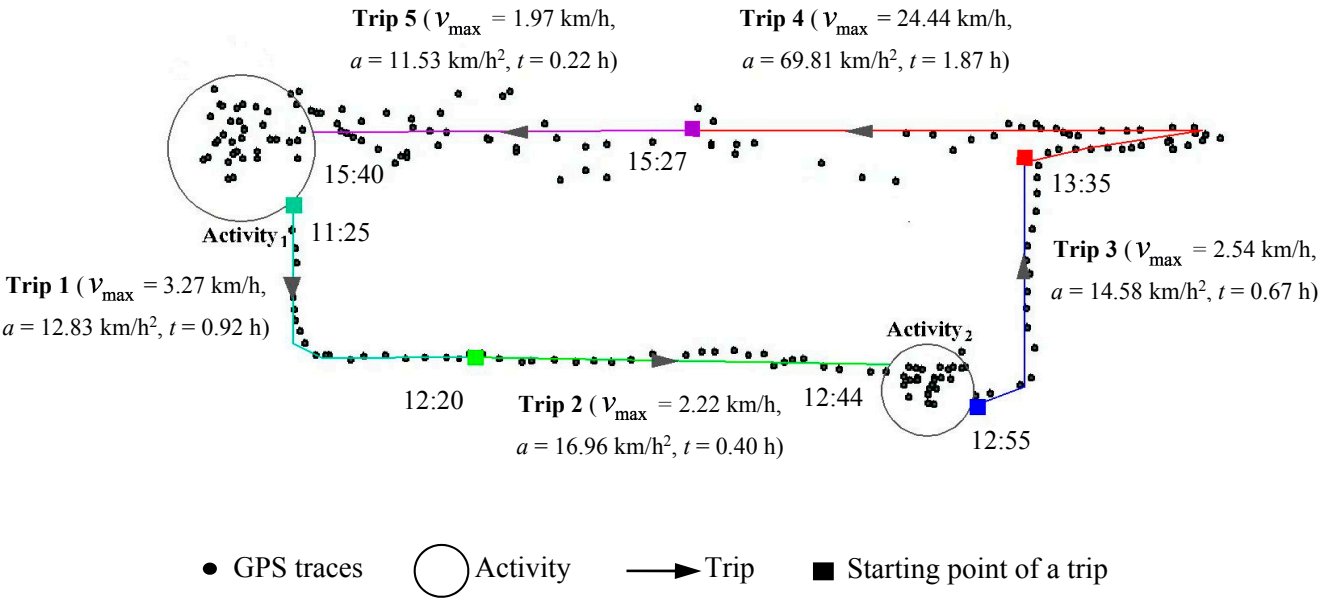

Figure 1. GPS traces and daily activity travel pattern of respondent #010.

4. Methodologies

In this section the major methods that will be used in mode detection are discussed. Being one type of SVM, SVC will be leveraged to identifying the traffic mode with GPS data. Compared to the multinomial logit (MNL) model, which has the limitation of determining alternatives with significant

correlation (such as bus and car), SVC can provide us with higher prediction accuracy. Besides, GA will also be employed to optimize the major parameters in SVC in order to enhance the solution quality and calculation efficiency of the model.

4.1. SVC

Being a popular disaggregate model, the MNL model has been widely used in mode prediction and identification. However, it imposes the restriction that the distribution of the random error terms is independent and identical over alternatives. This restriction leads to limitation of random taste variation and the independence of irrelevant alternatives, which causes the cross-elastics among all pairs of alternatives to be identical [21]. Compared to the MNL model, SVC has been more frequently applied in modeling disaggregate choices in recent years due to its ability to enhance the prediction accuracy and calculation efficiency [22]. Therefore, this paper will introduce SVC in estimating of travel mode with GPS data.

SVC is one of the SVM (*i.e.*, machine learning) methods analyzing data and recognizing patterns. Given a set of input-output data pairs (x_1, y_1), (x_2, y_2), …, (x_l, y_l) ($x_i \in X \subseteq R^m$, $y_i \in Y \subseteq R^n$, l is the number of training samples) that are randomly and independently generated from an unknown function, SVM estimates the function using the following Equation [23]:

$$f(x) = w \cdot \Phi(x) + b \qquad w, x \in R^m, b \in R^n, \tag{1}$$

where $\Phi(x)$ represents the high-dimensional feature spaces that are nonlinearly mapped from the input space x. w denotes a parameter vector and b is the threshold [24]. If the interpretation y only takes category values, *i.e.*, −1 and +1, it denotes SVC. Otherwise, if the domain of output space y contains continuous real values, the learning problem then refers to Support Vector Regression (SVR) [25].

For classification about the training data, SVM's linear soft margin algorithm is to solve the following regularized risk function:

$$MinJ = \frac{1}{2}\|w\|^2 + C \cdot R_{emp}[f] \tag{2}$$

The first term $\frac{1}{2}\|w\|^2$ is called the regularized term, which is used as a measurement of the function flatness. The second term $R_{emp}[f]$ is the so-called loss function to measure the empirical error. C is a regularization constant that determines the trade-off between the training error and the generalization performance. Here, the ε-insensitive loss function is employed to measure empirical error:

$$\left|y - f(x)\right|_\varepsilon = \max\{0, \left|y - f(x)\right| - \varepsilon\} \tag{3}$$

Equation (3) defines a ε tube (shown in Figure 2). The loss is zero if the predicted value is within the tube. If it is outside the tube, the loss is the magnitude of the difference between the predicted value and the radius ε of the tube. Both C and ε are user-determined parameters. Two positive slack variables ξ, ξ* are used to cope with the constraints of the optimization problem. To get the estimation of w and b, Equation (2) can be transformed to a primal objective function, Equation (4):

$$MinJ = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{s}(\xi_i^* + \xi_i)$$

$$s.t.\begin{cases} y_i - w\cdot\Phi(x_i) - b \le \varepsilon + \xi_i^* \\ w\cdot\Phi(x_i) + b - y_i \le \varepsilon + \xi_i \\ \xi_i^*, \xi_i \ge 0 \end{cases} \tag{4}$$

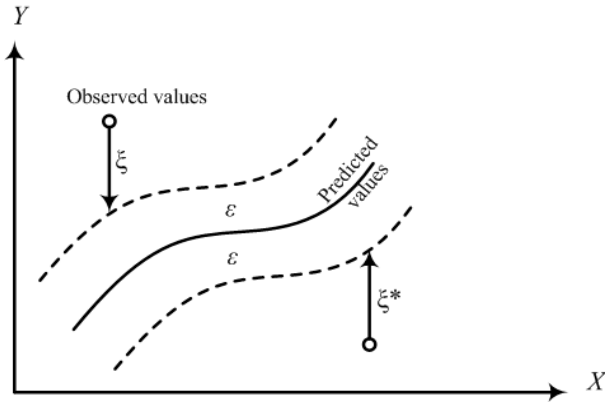

Figure 2. The parameters for SVC.

This constrained optimization problem is solved by using the following primal Lagrangian form:

$$\begin{aligned} L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{s}(\xi_i^* + \xi_i) - \sum_{i=1}^{s}(\eta_i\xi_i + \eta_i^*\xi_i^*) - \sum_{i=1}^{s}\alpha_i(\varepsilon + \xi_i - y_i + w\cdot\Phi(x_i) + b) \\ - \sum_{i=1}^{s}\alpha_i^*(\varepsilon + \xi^* - y_i - w\cdot\Phi(x_i) + b) \end{aligned} \tag{5}$$

where L is the Lagrangian, and η_i, η_i^*, α_i, α_i^* are Lagrange multipliers. Hence the dual variables in Equation (5) have to satisfy the positive constraints:

$$\eta_i, \eta_i^*, \alpha_i, \alpha_i^* \ge 0 \tag{6}$$

The above problem can be converted into a dual problem where the task is to optimize the Lagrangian multipliers, α_i and α_i^*. The dual problem contains a quadratic objective function of α_i and α_i^* with one linear constraint:

$$MaxJ = -\frac{1}{2}\sum_{i,j=1}^{s}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)(\Phi(x_i)\cdot\Phi(x_j)) + \sum_{i=1}^{s}\alpha_i^*(y_i - \varepsilon) - \sum_{i=1}^{s}\alpha_i(y_i + \varepsilon)$$

$$s.t.\begin{cases} \sum_{i=1}^{s}\alpha_i = \sum_{i=1}^{s}\alpha_i^* \\ 0 \le \alpha_i \le C \\ 0 \le \alpha_i^* \le C \end{cases} \tag{7}$$

Let

$$w - \sum_{i=1}^{s}(\alpha_i - \alpha_i^*)x_i = 0 \tag{8}$$

Thus,

$$f(x) = \sum_{i=1}^{s} (\alpha_i - \alpha_i^*) \Phi(x_i) \cdot \Phi(x_j) + b \tag{9}$$

By introducing kernel function $K(x_i,x_j)$, Equation (9) can be rewritten as follows:

$$f(x) = \sum_{i=1}^{s} (\alpha_i - \alpha_i^*) K(x_i, x_j) + b \tag{10}$$

where $K(x_i, x_j)$ is the so-called kernel function which is equal to the inner product of two vectors x_i and x_j in the feature space $\Phi(x_i)$ and $\Phi(x_j)$; that is, $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$.

The kernel function would be more simply without the compution of Φ(X). Some popular kernel functions are the linear kernel, polynomial kernel, and radial-basis function (RBF) kernel. Using different kernel functions, one can construct different learning machines with arbitrary types of decision surfaces. In general, the RBF kernel, as a nonlinearly kernel function, is a reasonable first choice [26]. Thus, the RBF kernel is chosen in this work:

$$K(x_i, x_j) = \exp\left(-\frac{\left\| x_i - x_j \right\|^2}{2\sigma^2}\right) \tag{11}$$

where σ is a parameter that determines the area of influence this support vector has over the data space.

As user-determined parameters in SVC and the RBF kernel, C, ε, and σ will greatly influence the estimation efficiency and prediction accuracy of the models, especially for large-scale or real-time feature practice application. This paper will optimize the parameters by employing GA. Being a part of evolutionary computing, GA is a rapidly growing area of artificial intelligence. The process of GA is presented as follows.

4.2. Parameter Optimization with GA

Being the key elements in SVC, the parameters C, ε, and σ directly decide the prediction performance of the model. Therefore, the parameter optimization is an important factor for improving the prediction accuracy of SVC. In this paper, GA is applied to optimize C, ε, and σ in SVC. GA is inspired by evolutionary biology processes like inheritance, selection, crossover, and mutation. Based on a fitness function, GA attempts to retain relatively good genetic information from generation to generation. The process of GA can be divided into six steps, which will be described briefly in the following section.

4.2.1. Encoding of Chromosomes

GA starts with a set of solutions (represented by chromosomes) called population. The individuals comprising the population are known as chromosomes. In most GA applications, the chromosomes are encoded as a series of zeroes and ones, or a binary bit string. For the mode detection model with SVC, the real encodings were adopted since the parameters C, ε, and σ are continuous-valued. To represent the parameters in SVC, each chromosome consists of gen_1^n, gen_2^n, and gen_3^n (n refers to the current generation), which represent three parameters, respectively.

To reduce the search space referring to previous literature using SVC, the three parameters should within the range $C \in [2^{-6}, 2^{6}]$, $\varepsilon \in [2^{-12}, 2^{-1}]$, and $\sigma \in [0, 2]$ [27]. An example of the encoding of a chromosome is shown in Table 2.

Table 2. An example of chromosome encoding.

$2^{-6} \le C \le 2^{6}$	$2^{-12} \le \varepsilon \le 2^{-1}$	$0 \le \sigma \le 2$
$gen^{n}{}_{1}$	$gen^{n}{}_{2}$	$gen^{n}{}_{3}$

4.2.2. Fitness Function

Fitness function determines possible solutions to the problem and is used to estimate the quality of the represented solution (chromosome). For parameter optimizations in SVC, the best solution is able to maximize the accuracy rate of prediction. Generally, GA is an optimal searching method to find the maximum fitness of the individual chromosome. Thus, Hit ratio is adopted in this paper. Here, Hit ratio (*HitR*) refers to the fitting degree of the identification results to the observed samples.

$$fit = \frac{N_1}{N_2} \tag{12}$$

where N_1 is the number of hit records (travel mode) predicted by the model and N_2 is the total number of observations.

4.2.3. Crossover Operator

Crossover is a reproduction technique that takes two parent chromosomes and produces two child chromosomes. In this paper, an arithmetic crossover is used to create new offspring [28].

$$\begin{aligned} gen_{k,I}^{n} &= \alpha_i gen_{k,I}^{n-1} + (1-\alpha_k) gen_{k,II}^{n-1} \\ gen_{k,II}^{n} &= \alpha_i gen_{k,II}^{n-1} + (1-\alpha_k) gen_{k,I}^{n-1} \end{aligned} \tag{13}$$

where $gen_{k,I}^{n-1}$, $gen_{k,II}^{n-1}$ is a pair of "parent" chromosomes; $gen_{k,I}^{n}$, $gen_{k,II}^{n}$ is a pair of "children" chromosomes; α_k is a random number between (0, 1); and $k \in [1,2,3]$ (k is the total genes of the crossover operation). Table 3 shows the parents selected for crossover. When k = 1 and α_k = 0.4, the children chromosomes after crossover are shown in Table 4.

Table 3. An example of chromosome encoding.

Parent I	$gen^{n-1}{}_{1,I}$	$gen^{n-1}{}_{2,I}$	$gen^{n-1}{}_{3,I}$
Parent II	$gen^{n-1}{}_{1,II}$	$gen^{n-1}{}_{2,II}$	$gen^{n-1}{}_{3,II}$

Table 4. Children chromosomes after crossover.

Children I	$0.4gen^{n}{}_{1,I} + 0.6gen^{n}{}_{1,II}$	$gen^{n-1}{}_{2,I}$	$gen^{n-1}{}_{3,I}$
Children II	$0.4gen^{n}{}_{1,II} + 0.6gen^{n}{}_{1,I}$	$gen^{n-1}{}_{2,II}$	$gen^{n-1}{}_{3,II}$

4.2.4. Mutation Operator

Mutation is a common reproduction operator used for finding new points in the searching space to evaluate. A genetic mutation operation is used in this paper [27].

Assuming a chromosome is $G = (gen_1^n, gen_2^n, gen_3^n)$, if the gen_1^n is selected for the mutation, the mutation can be shown with Equation (14):

$$G' = (gen_1^{n-1}, gen_2^{n-1}, gen_3^{n-1})$$
$$gen_1^n = \begin{cases} gen_1^{n-1} + \Delta(n, gen_{1,\max}^n - gen_1^{n-1}) & if \quad random(0,1) = 0 \\ gen_1^{n-1} + \Delta(n, gen_1^{n-1} - gen_{1,\min}^n) & if \quad random(0,1) = 1 \end{cases} \tag{14}$$

Then, the function $\Delta(n, y)$ returns a value between [0, y] given in Equation (15):

$$\Delta(n, y) = y \times (1 - r^{(1-n/T_{\max})\lambda}) \tag{15}$$

where r is a random number between [0,1]; $T_{\max}$ is maximum number of generations; and λ = 3. This property causes this operation to make a uniform search in the initial space when n is small, and a very local one in later stages.

To deal with the problem that the mutation may violate the parameters' constraints, this paper will assign a relatively high weight to reduce their probability of being selected in the following search [27,29].

4.2.5. Termination

The search continues until $HitR_n$-$HitR_{n-1}$ < 0.001% or the number of generation reaches the maximum number of generations $T_{\max}$, which is set to be 5000 [30].

4.2.6. The procedure of GA

The major steps of GA are shown in Figure 3.

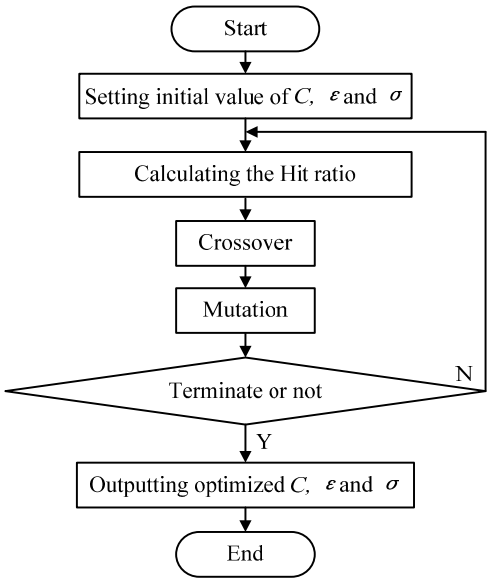

Figure 3. The basic procedure of GA.

5. Identification Model

In this section the detailed setting of parameters in SVC and GA as well as the modeling procedure of mode detection are introduced. The models are estimated with the survey data in Beijing. Compared to the observed records provided by the survey data, success rates of 100%, 100%, 88.9%, 92.7%, and 80.0% were obtained for the detection of walking, bicycle, subway, bus, and car, respectively. This indicates that the developed model shows a high level of accuracy rate for mode identification. For a case study, the identified and observed mode choices of respondent #010 are presented.

5.1. Alternatives

According to the travel survey data, the real mode share in Beijing in 2010 was calculated. The results are shown in Table 5. The results reveal that car and bus are the two most popular travel modes in Beijing. For mode identification using GPS data, buses and cars have great correlation due to their similar characteristics, especially concerning the speed and acceleration in almost the same traffic conditions. Therefore, as mentioned above, SVC, instead of the MNL model, will be employed to establish the mode identification model in this paper [31,32].

Table 5. Mode share in Beijing, 2010.

Mode	Walking	Bicycle	Subway	Bus	Car	Other
Percentage (%)	10.7	15	6.9	22.2	30	15.2

5.2. Variables

Based on a preliminary correlation test, eight variables related to the characteristics of travel modes were selected, as shown in Table 6.

Table 6. Variables in the mode identification model.

Variables		Mean [1]	S.D. [2]	Variables		Mean	S.D.
Average Speed (km/h)	v_{walk}	3.00	2.49	*Average acceleration* (km/h^2)	a_{walk}	57.97	43.26
	$v_{bicycle}$	8.47	8.20		$a_{bicycle}$	60.90	80.81
	v_{subway}	11.21	13.32		a_{subway}	73.47	118.94
	v_{bus}	13.27	14.21		a_{bus}	78.99	115.95
	v_{car}	14.54	9.20		a_{car}	84.58	103.59
Maximum speed (km/h)	$v_{max\text{-}walk}$	15.35	7.45	*75th percentile of acceleration* (km/h^2)	$a_{75th\%\text{-}walk}$	434.20	400.86
	$v_{max\text{-}bicycle}$	20.87	14.73		$a_{75th\%\text{-}bicycle}$	864.85	431.41
	$v_{max\text{-}subway}$	44.51	30.22		$a_{75th\%\text{-}subway}$	722.90	754.76
	$v_{max\text{-}bus}$	50.23	19.33		$a_{75th\%\text{-}bus}$	2012.55	1203.71
	$v_{max\text{-}car}$	53.25	18.73		$a_{75th\%\text{-}car}$	2290.54	1327.73

Table 6. *Cont.*

Variables		Mean [1]	S.D. [2]	Variables		Mean	S.D.
75th percentile of speed (km/h)	$v_{75th\%\text{-}walk}$	3.62	2.01	*Travel time* (h)	t_{walk}	2.56	3.17
	$v_{75th\%\text{-}bicycle}$	12.44	10.47		$t_{bicycle}$	3.46	4.92
	$v_{75th\%\text{-}subway}$	24.19	31.92		t_{subway}	0.56	0.30
	$v_{75th\%\text{-}bus}$	25.15	12.32		t_{bus}	1.21	1.12
	$v_{75th\%\text{-}car}$	24.48	15.47		t_{car}	0.64	0.73
Standard deviation of speed	$\sigma_{walking}$	4.80	7.68	*Trip distance* (km)	d_{walk}	0.36	0.29
	$\sigma_{bicycle}$	0.92	0.52		$d_{bicycle}$	2.03	2.76
	σ_{subway}	2.53	1.80		d_{subway}	4.52	4.96
	σ_{bus}	2.37	2.54		d_{bus}	7.59	9.26
	σ_{car}	1.62	3.44		d_{car}	8.17	7.48

Note: [1] Mean denotes mean value; [2] S.D. refers to standard deviation.

5.3. Estimation Results

There are three major GA parameters, namely p_c, p_m, and p_{size}. In general, p_c varies from 0.3 to 0.9 and p_m varies from 0.01 to 0.1, while p_{size} is the population size, which is set according to the size of the samples. Considering the features of mode identification, the condition of survey data, and the previous studies [27,28,33,34] related to GA, p_c, p_m, and p_{size} are set at 0.6, 0.06, and 80, respectively. The convergence of the calculation by GA indicates that the Hit ratio increases slowly after the 3000th generation. The highest Hit ratio appears in about the 3500th generation, and remains almost unchanged after that. The three parameters, *i.e.*, C, ε, and σ were then optimized as 84.54, 0.0009, and 0.7367, respectively, with the best optimization value among the 10 results for the practical mode detection model.

The estimation results of the SVC model are shown in Table 7. The results indicate that the factors related to acceleration, travel speed, and travel time are the major ones that should be considered in travel mode detection, while that regarding trip distance does not have significant impact on mode decision. One of the reasons for this is that there is correlation between travel time and trip distance. In detecting walking trips, travel time and maximum speed are important factors—that is, the longer a trip takes or the larger its maximum speed is, there is less probability that its mode is walking. The results also show that there are three major factors, *i.e.*, the 75th percentile of speed, maximum speed, and acceleration, which impact the detection of a bicycle. In detail, a trip with low speed and acceleration tends to be by bicycle. On the contrary, a trip with high speed and acceleration is more likely to be on the subway. The determination of a bus is similar to that of subway, except that the coefficient of maximum speed and acceleration is smaller than that of the subway. This reveals that the higher a trip's speed and acceleration, the more probability that it refers to the mode of subway.

Table 7. Estimation results of the mode detection model.

Variables	Coefficient	Standard Error	T-statistic
$t_{walking}$	−6.009	3.375	−1.78
$v_{max\text{-}walking}$	−0.861	0.394	−2.19
$v_{75th\%\text{-}bicycle}$	−0.351	0.172	−2.04
$a_{bicycle}$	−0.281	0.120	−2.34
$v_{max\text{-}bicycle}$	−0.412	0.157	−2.62
$a_{75th\%\text{-}subway}$	−0.004	0.001	−4.00
a_{subway}	0.172	0.049	3.51
$v_{max\text{-}subway}$	0.474	0.169	2.80
$v_{max\text{-}bus}$	0.006	0.002	3.00
a_{bus}	0.061	0.015	4.07

Note: Dependent variable = mode; Number of sample = 85,120.

5.4. Mode Identification

The travel mode of each trip in the sample data is determined by using the developed model. Figure 4 shows respondent #010's daily mode choices identified by the model.

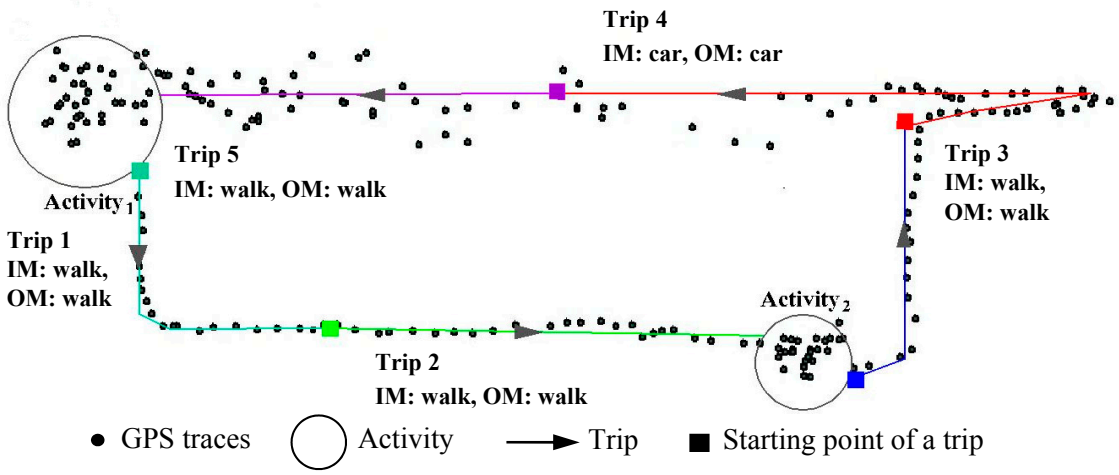

Figure 4. Identified and observed mode choices of respondent #010.

With the developed model, the mode choices of all the trips in the GPS data are determined. By comparing the predicted results and the observed records, the Hit ratios of the model are shown in Table 8. The estimation results show high accuracy for detection of walking, bicycle, subway, and bus, and that of car is also acceptable. Further analysis reveals that the reason for the relatively lower accuracy of car and subway determination is that the model sometimes makes mistakes in distinguishing between car and subway. Furthermore, compared to the detection results presented by Gong *et al.* [3]—*i.e.*, 92.4%, 65.5%, 62.5%, and 84.1% for walking, subway, bus, and car, respectively—the detection results of this work are better. This also represented the good performance of the SVC model in mode detection.

Table 8. Verification results of the mode detection model.

	Identified by algorithm as					Total trip segments	Correct segments	Success rate (%)
	Walking	**Bicycle**	**Subway**	**Bus**	**Car**			
Walking	24,961	0	0	0	0	24,961	24,961	100.0
Bicycle	0	12,685	0	0	0	12,685	12,685	100.0
Subway	0	0	3250	204	202	3656	3250	88.9
Bus	206	409	0	18,209	818	19,642	18,209	92.7
Car	129	205	2864	1637	19,341	24,176	19,341	80.0
Total	25,296	13,299	6114	20,050	20,361	85,120	78,446	92.2

6. Conclusions

In this paper, an SVC model was constructed for mode detection with GPS survey data. GA was used in optimizing the parameters in SVC. The results indicate that the developed model shows a high level of accuracy for mode identification. Our findings can significantly enhance the efficiency and accuracy of travel survey and data processing. They also serve as a foundation for a future model system of full-scale travel information identification with GPS data. Moreover, by providing crucial travel information, the results contribute to the modeling and analyzing of travel behavior and are readily applicable to a wide range of transportation practice.

The estimation results also reveal that further study should be conducted with respect to high-accuracy detection of subway, bus, and car. One of the potential methods is to distinguish subway and bus from car based on the GIS information. Therefore, a future study could make potential progress by combining GPS data with GIS technology to determine travel modes. Moreover, although the sample size is relatively large compared to some of previous studies, it was not representative enough for the general population. Further study should pay more attention to the process of sampling and data collection to eliminate sampling bias.

Acknowledgments

The research was funded by the National Natural Science Foundation of China (50908099), the Humanity and Social Science Youth Foundation of the Ministry of Education (14YJC630225), the China Postdoctoral Science Special Foundation (2014M551191), and Jilin University's Outstanding Youth fund (2013JQ007). Special thanks to the anonymous reviewers whose valuable comments helped tremendously in improving this paper.

Author Contributions

The research scheme was mainly designed by Fang Zong and Yu Bai. Xiao Wang, Yixin Yuan and Yanan He performed the research and analyzed the data. The paper was mainly written by Fang Zong. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Stopher, P.; Greaves, S.P. Household travel surveys: Where are we going? *Transport Res A*. **2007**, *41*, 367–381.
2. Stopher, P.R.; Metcalf, H.M.A. *Metcalf, Synthesis of Highway Practice No. 236: Methods for Household Travel Surveys*; NA Press: Washington, DC, USA, 1996.
3. Gong, H.; Chen, C.; Bialostozky, E.; Lawson, C.T. A GPS/GIS method for travel mode detection in New York City. *Comput. Environ. Urban Syst.* **2012**, *36*, 131–139.
4. Liu, J.; Zheng, H.J. Post-processing procedures for passive GPS based travel survey. *Procedia Soc. Behav. Sci*. **2013**, *96*, 310–319.
5. Bohte, W.; Maat, K. Deriving and validating trip destinations and modes for multi-day GPS based travel surveys: An application in the Netherlands. *Transp. Res. C* **2009**, *17*, 285–297.
6. Byon, Y.J.; Abdulhai, B.; Shalaby, A.S. Impact of sampling rate of GPS-enabled cell phones on mode detectionand GIS map matching performance. In Proceedings of the Transportation Research Board 86th Annual Meeting, Washington, DC, USA, 21–25 January 2007.
7. Gonzalez, P.; Weinstein, J.; Barbeau, S. Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones. In Proceedings of the 15th World Congress on ITS, New York, NY, USA, 10–15 January 2008; pp. 297–300.
8. Patterson, D.J.; Liao, L.; Fox, D.; Kautz, H. Inferring high-level behavior from low-level sensors. In *UbiComp 2003: Ubiquitous Computing*; Anind, K.D., Albrecht, S., Joseph, F.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2003; pp. 73–89.
9. Zheng, Y.; Liu, L.; Wang, L.; Xie, X. Learning transportation mode from raw GPS data for geographic applications on the web. In Proceedings of the WWW 2008, Beijing, China, 21–25 April 2008; pp. 247–256.
10. Reddy, S.; Burke, J.; Estrin, D.; Hansen, M.; Srivastava, M. Determining transportation mode on mobile phones. In Proceedings of the 12th IEEE International Symposium on Wearable Computers, Pittsburgh, PA, USA, 28 September–1 October 2008; pp. 25–28.
11. Moiseeva, A.; Timmermans, H. Imputing relevant information from multi-day GPS tracers for retail planning and management using data fusion and context-sensitive learning. *J. Retail. Consum. Serv.* **2010**, *17*, 189–199.
12. Feng, T.; Timmermans, H.J.P. Transportation mode recognition using GPS and accelerometer data. *Transp. Res. C* **2013**, *37*, 118–130.
13. Zhang, L.; Dalyot, S.; Eggert, D.; Sester, M. Multi-stage approach to travel-mode segmentation and classification of GPS traces. In Proceedings of the ISPRS Guilin 2011 Workshop on International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Guilin, China, 20–21 October 2011; pp. 87–93.
14. Pereira, F.; Carrion, C.; Zhao, F.; Cottrill, C.D.; Zegras, C.; Ben-Akiva, M. The future mobility survey: Overview and preliminary evaluation. In Proceedings of the 10th International Conference of Eastern Asia Society for Transportation Studies, Taipei, Taiwan, 9–12 September 2013; p. 31.
15. Zhang, Z.H. Deriving Trip Information from GPS Trajectories. Ph.D. Thesis, East China Normal University, Shanghai, China, June 2010. (In Chinese)
16. Vij, A.; Carrel, A.; Walker, J.L. Incorporating the influence of latent modal preferences on travel

mode choice behavior. *Transp. Res. A* **2013**, *54*, 164–178.
17. Du, J.; Aultman-Hall, L. Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues. *Transp. Res. A* **2007**, *41*, 220–232.
18. Bolbol, A.; Cheng,T.; Tsapakis, I. A spatio-temporal approach for identifying the sample size for transport mode detection from GPS-based travel surveys: A case study of London's road network. *Transp. Res. C* **2014**, *43*, 176–187.
19. Zong, F.; Wang, X.; Zhang, H.Y.; Bai, Y. Identifying trip-activity-intermediate stop using GPS-based travel survey data. *J. SCUT. (Nat. Sci. Ed.)*. **2015**, *43*, 28–32. (In Chinese)
20. Sehonfelder, S.; Samaga, U. Where do You Want to Go Today?—More Observations on Daily Mobility. In Proceedings of the 3rd Swiss Transport Research Conference (STRC), Ascona, Switzerland, 19–21 March 2003; Available online: http://www.strc.ch/Paper/Schoen.pdf (accessed on 28 November 2004).
21. Wen, C.H.; Koppelman, F.S. The generalized nested logit model. *Transp. Res. B* **2001**, *35*, 627–641.
22. Zong, F.; Jia, H.F.; Pan, X.; Wu, Y. Prediction of commuter's daily time allocation. *PROMET Traffic Transp.* **2013**, *25*, 445–455.
23. Sung, K.; Poggio, T. Example-based learning for view-based human face detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 39–50.
24. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer: New York, NY, USA, 1995.
25. Li, Y.M.; Gong, S.G.; Liddell, H.M. Support Vector Regression and Classification based multi-view face detection and recognition. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, 28–30 March 2000; pp. 300–305.
26. Dong, B.; Cao, C.; Lee, S.E. Applying Support Vector Machines to predict building energy consumption in tropical region. *Energ. Build*. **2005**, *37*, 545–553.
27. Yao, J.B.; Yao, B.Z.; Li, L.; Jiang, Y.L. Hybrid model for displacement prediction of tunnel surrounding rock. *Neural Netw. World* **2012**, *22*, 263–275.
28. Yu, B.; Yang, Z.Z.; Cheng, C. Optimizing the distribution of shopping centers with parallel genetic algorithm. *Eng. Appl. Artif. Intell*. **2007**, *20*, 215–223.
29. Yao, B.Z.; Hu, P.; Lu, X.H.; Gao, J.J.; Zhang, M.H. Transit network design based on travel time reliability. *Transp. Res. C* **2014**, *43*, 233–248.
30. Zong, F.; Lin, H.Y.; Yu, B.; Pan, X. Daily commute time prediction based on genetic algorithm. *Math. Probl. Eng*. **2012**, doi:10.1155/2012/321574.
31. Este, A.; Gringoli, F.; Salgarelli, L. Support Vector Machines for TCP traffic classification. *Comput. Netw*. **2009**, *53*, 2476–2490.
32. Anguita, D.; Boni, A.; Ridella, S. Evaluating the generalization ability of Support Vector Machines through the bootstrap. *Neural Process. Lett.* **2000**, *11*, 51–58.
33. Wang, Y.Q.; Li, Y.; Wang, Q.; Lv, Y.L.; Wang, S.Y.; Chen, X.; Yu, X.X.; Jiang, W.; Li, X. Computational identification of human long intergenic non-coding RNAs using a GA-SVM algorithm. *Gene* **2014**, *533*, 94–99.
34. Lee, Y.; Lee, J. Binary tree optimization using genetic algorithm for multiclass support vector machine. *Expert Syst. Appl*. **2015**, *42*, 3843–3851.