Categorizing three active cyclist typologies by exploring patterns on a multitude of GPS crowdsourced data attributes

Published Version:
Poliziani C., Rupi F., Mbuga F., Schweizer J., Tortora C. (2021). Categorizing three active cyclist typologies by exploring patterns on a multitude of GPS crowdsourced data attributes. RESEARCH IN TRANSPORTATION BUSINESS & MANAGEMENT, 40, 1-8 [10.1016/j.rtbm.2020.100572].

Availability:
This version is available at: https://hdl.handle.net/11585/796086 since: 2023-01-30

Published:
DOI: http://doi.org/10.1016/j.rtbm.2020.100572

RESEARCH IN TRANSPORTATION BUSINESS &MANAGEMENT

Categorizing three active cyclist typologies by exploring patterns on a multitude of GPS crowdsourced data attributes.

Paper for peer review. Special issue: NECTAR2020, Venice

Cristian Poliziani[1], Federico Rupi[1], Felix Mbuga[2], Joerg Schweizer[1], Cristina Tortora[2]

[1]DICAM- University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy

[2]San José State University, One Washington square, 95192 San Jose, CA USA

Abstract

The paper tackles the problem of characterizing cyclist typologies using a data set of GPS traces. The data is crowdsourced and consists of 29,431 traces recorded in the city of Bologna, Italy during the morning rush hours from 7am to 10am of work-days and during the months from April through September 2017. Different criteria to group the heterogenous behavior of cyclists into separate categories have already been described in literature, where studies are generally based on stated preference interviews or on observing a small sample of the population. The novelty of this study is clustering a data set based on GPS traces from 2,135 cyclists, which is the equivalent to a revealed survey. Furthermore, refined pre-processing of the GPS traces allows the determination of dynamic attributes, a comparison of the chosen path respect to the shortest path and the evaluation of other specific trip attributes, which are either difficult or impossible to assess by a classical interview. The applied clusterization process leads to 3 main cyclist typologies, where each type is characterized by different trip attributes and behaviors involving safety, riskiness, precaution, inexperience, knowledge, fear and hastiness: *Risky & Hasty*, *Inexperienced & Inefficient and Smart & Informed*.

Keywords

Cyclist typology, Cluster analysis, GPS trace, Big data

Highlights

- *Crowdsourced big data of GPS traces for the characterization of cyclist typologies*
- *Cluster analysis with more than 50 trip characterizing attributes.*
- *Three cyclist typologies: Risky & Hasty, Inexperienced & Inefficient and Smart & Informed*

1. Introduction

1.1 Motivation and scope

The availability of a big set of GPS traces from cyclists, together with a range of instruments and tools able to deeply analyze GPS data has led to the idea of identifying different types of cyclists using a cluster analysis method, which is the core of the present work. The motivation behind this approach is to replace interview-based surveys with a large quantity of revealed data in order to obtain a more objective classification of cyclists. The knowledge of the type of cyclist is relevant for the link-cost and route-choice modeling of cyclists.

1.2 Literature review and State of the art

In order to properly plan bicycle infrastructure, it is crucial to know and characterize the current transportation demand. However, data collection for the purpose of demand modeling is a challenge faced by many cities. Most municipalities do not systematically monitor cycling activities.

Cyclists are heterogeneous and show different riding behavior. The differentiation between individual user groups by targeting more accurately the needs and requirements of different types of users aims to better conduct bicycle infrastructure planning, model cyclist behaviors, identify critical points and reduce barriers for cycling. Planning a network adapted to different cyclist types could be an effective strategy to increase cycling mode share and frequency among the various groups (Damant-Sirois et al., 2014).

Many studies are based on survey data to group them according to different aspects of their riding behavior. The main cyclist typology factors found as trip purpose (e.g., Kroesen and Handy, 2014), comfort/perception of safety (Geller, 2006) as well as motivational (e.g., Gatersleben and Haddad, 2010; Damant-Sirois et al., 2014) and social factors such as identification as a cyclist (Damant-Sirois et al., 2014). Kroesen and Handy (2014) considered non cyclists and cyclists for work and non-work purposes and used a latent transition model for clustering people into four groups: non-cyclists, non-work cyclists, all-around cyclists (for work and non-work purposes) and commuter cyclists. Gatersleben and Haddad (2010), using the results of a survey conducted amongst 244 cyclists and non-cyclists in England, distinguished four different bicyclist types based on behavior, motivation and characteristics of the typical bicyclist: responsible, lifestyle, commuter and day-to-day. These types differed between bicyclists and non-bicyclists.

Geller (2006) identified four types of cyclists, subjectively developed on the basis on his expert knowledge: strong and fearless, enthused and confident, interested but concerned, and no way no how. This typology is based on perceived safety (comfort level on different types of bikeways and fear of people driving automobiles) and on people's interest in cycling more. This classification, referred to all adults regardless of their current cycling behavior, was subsequently formalized in a method and validated by Dill and McNeil by a random phone survey in a sample of adults in the 50 largest metro regions in the U.S. (Dill and McNeil, 2013; 2016). Damant-Sirois et al. (2014), using data from an online survey aimed only at cyclists, propose a multidimensional cyclist typology based on seven factors including weather conditions and effort, time efficiency, street design, bicycle facilities, personal identity toward cycling and past cycling history. They distinguished four distinct cyclist types: dedicated cyclists, path-using cyclists, fairweather utilitarian and leisure cyclists. More recently, Cabral and Kim (2020) question the classic Four Types of Cyclists proposed by Geller, particularly with respect to perceived comfort. They used an online survey and video clips to classify people into three categories: Uncomfortable or Uninterested, Cautious Majority, and Very Comfortable Cyclists. Their empirical segmentation is based on variables of comfort, cycling intent, and cycling in the previous summer.

Francke et al. (2020) propose a multidimensional typology of cyclists which includes the influence factors of already existing studies complemented by motivational factors. They use an empirical approach through a Germany-wide online survey (10,294 responses) on cycling behavior in order to distinguish four distinct types of cyclists: ambitious, functional, pragmatic, and passionate cyclists.

The use of recorded GPS data instead of interviews may be the way forward as GPS data represent objective information about the chosen route and motion of each cyclist. A large number of GPS traces are usually available from bike-tracing campaigns. In recent years, many studies on cycling mobility have made use of GPS data, which is often available at low cost and allows to gain a broad range of information, such as the spatial distribution of cyclists on the city's road network. Such information allows to calibrate the cyclist route choice model, see (Lu et al., 2018; Rupi et al., 2019; Rupi and Schweizer, 2018; Pritchard et al., 2019; Pritchard , 2018; Griffin and Jiao, 2015; Charlton et al., 2011; Dill, 2009; Menghini et al., 2010; Hood et al., 2011; Broach et al., 2012; Zimmermann et al., 2017; Bernardi et al., 2018; Schweizer et al., 2020; Chen et al., 2018). Other studies use GPS traces to obtain information on cyclist speeds (Manum et al., 2018, Flügel et al., 2017, Strauss et al., 2017), speed profiles (Strauss and Miranda-Moreno, 2017; Clarry et al., 2019; Laranjeiro et al., 2019) and waiting times at intersections (Watkins and LeDantec, 2016; Rupi et al. 2020).

1.3 Research contribution

The present study explores a new method to identify and characterize different types of cyclists based on GPS traces and some additional attributes of cyclists. The method is applied to the GPS traces recorded in the city of Bologna, Italy. The novelty is the use of GPS traces and derived quantities (such as trip-length, speeds, waiting times, deviation from shortest path, etc.) instead of interviews for the purpose of identifying types of cyclist, which is equivalent to a revealed preference survey – hence the GPS traces are expected to be more reliable than classical surveys, where interviewees declare their behavior. In addition, GPS traces are often available in large quantities. The described method requires a refined pre-processing of the GPS traces: The traces are matched to a road network extracted from Open Street Map (OSM) and elaborated in order to create a rich database that describes the experience, the performance, choices and behavior of each cyclist as detailed in Rupi et al., 2020.

The main focus of the paper is on the successive cluster analysis which uses the results from the pre-processing step with the scope of clustering the data set in groups of cyclists characterized by similar habits.

Section 2 describes the applied clustering method and section 3 presents the Bologna GPS data set and briefly explains the data pre-processing. The results in section 4 present the 3 types of cyclists and illustrate their respective characteristics.

2. Cluster analysis

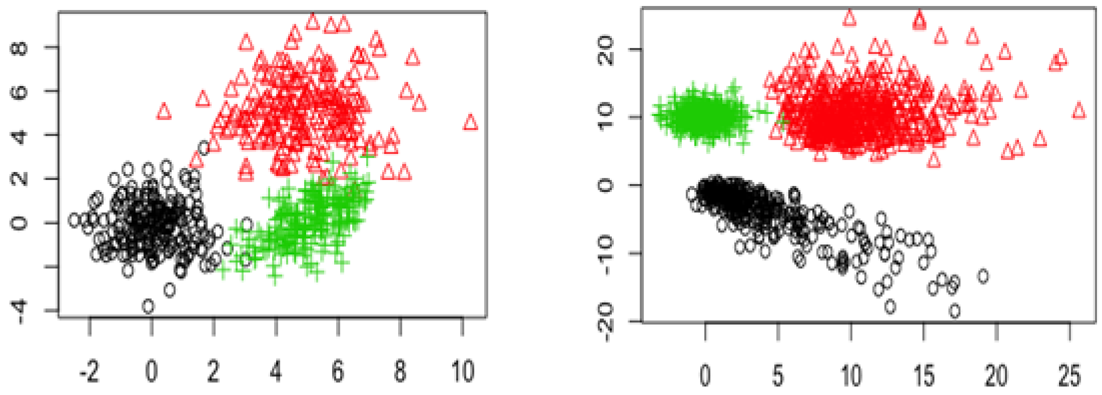
The goal of cluster analysis is to find homogeneous groups of units within the data, i.e. homogeneous groups of cyclists based on their habits. There are many clustering techniques in the statistical literature (see Everitt 2011 for examples), among them model based clustering techniques are popular. Model-based clustering assumes that a population is a convex combination of a finite number of density functions. The multivariate Gaussian distribution is one of the most popular for its simplicity (McLachlan and Basford 1988): each cluster has only two parameters, the mean vector that determines the position of the cluster in the space, and the covariance matrix.

Figure 1 (a) shows an example of a two-dimensional data set with 3 clusters that follow a bivariate Gaussian distribution. In cluster one (black circles) the variables have unitary variance and no correlation, in cluster two (red triangles) the variables have variance equal to 2 and no correlation, and in cluster three (green plues) the variables have unitary variance and correlation equal to 0.6.

Formally, a p-dimensional random vector X follows a finite mixture of distributions if, for all $x \subset X$, its density can be written as: $f(x|\vartheta)=\sum_{g=1}^{G}\pi_g f(x|\theta_g)$, where G is the number of clusters, $\pi_g>0$, such that $\sum_{g=1}^{G}\pi_g=1$, is the gth mixing proportion, $f(x|\theta_g)$ is the gth component density, and $\vartheta=(\pi_1,\dots,\pi_G,\theta_1,\dots,\theta_G)$ is the vector of parameters. One of the advantages of model-based clustering is that, besides the partition in clusters, the method produces an estimate of the parameters θ_g of each cluster that can be used for interpretation.

A variety of distributions have been used to model the density functions; McNicholas 2016 contains a good review of them. Among these distributions, the generalized hyperbolic distribution (GHD) (Browne and

McNicholas 2015) has the advantage of being extremely flexible. The GHD is characterized by five parameters – mean vector, scale matrix, skewness vector, concentration, and index parameters - and can identify clusters of a different shape. Moreover, many other distributions - like the Gaussian distribution, the multivariate Student t, or the skewed Student t distribution - can be obtained as a special or a limiting case of the GHD, i.e. it can detect clusters that are for example normally distributed. Tortora et al. 2019 proposed a flexible extension of the GHD - the coalesced GHD (CGHD)- that adds even more flexibility and can model non-convex clusters. Figure 1 (b) shows an example of a two-dimensional data set with 3 clusters that follow a bivariate CGHD. The Bayesian information criterion (BIC) is recommended to choose the number of clusters for mixtures of CGHD.



(a) (b)

Fig.1 Two dimensional data sets where each cluster was generated from a Gaussian distribution (a) or from a CGHD (b).

3. The Bologna data set

The city of Bologna, Italy, hosted in 2017 - from April 1st to September 30th - the 'Bella Mossa' initiative (BM), funded by the EU and Bologna's local Government. The initiative had the objective to promote sustainable mobility by rewarding people (with coupons for local shops) for recording their GPS traces of sustainable trips - meaning trips done by transit, bike or walking . The smartphone application 'Betterpoints' (Betterpoints, 2020) has been used to record and collect the data. The full data sample contains approximately 270,000 bike GPS traces, which consist of more than 62 million points: the smartphone application records one GPS point every 5 seconds when the bike is in motion. When the bike stops, for example at intersections, the recording stops also. The present study focuses only on bikes GPS traces recorded during the morning peak of work days - from 7am to 10am - because of computational time problems, but also trying to englobe especially working trips during the morning peak hour and avoid as much as possible trips for other purposes.

The following data processing steps have been implemented using the SUMOPy environment (Schweizer, 2020), an open source extension of the software SUMO (Eclipse SUMO, 2020). In a first step, the open street map (OSM) network covering the urban area of Bologna (OpenStreetMap, 2020) has been imported into SUMO. This SUMO network is attribute-rich and contains information on road width, road access (e.g. reserved bikeways, shared access, with pedestrians, etc.) and speed-limits. From these basic attributes SUMO derives a road priority (1-10), where low priority roads are taken values from 1 to 7. Successively the network has been manually improved in order to eliminate errors due to an imperfect OSM representation as well as conversion errors. Next, unrealistic GPS traces have been deleted: trips outside the study area and traces which have probably not been recorded while riding a bike. (See Fig.2). Valid traces must also satisfy criteria such as a certain total distance, a minimum number of points, and a minimum and maximum distance between successive points. This trace filtering step does ensure the GPS traces can be successfully matched to the road network by the map-matching process. During the map-matching the most likely route (as sequence of network links) can be identified for each GPS trace (Schweizer et al., 2020).

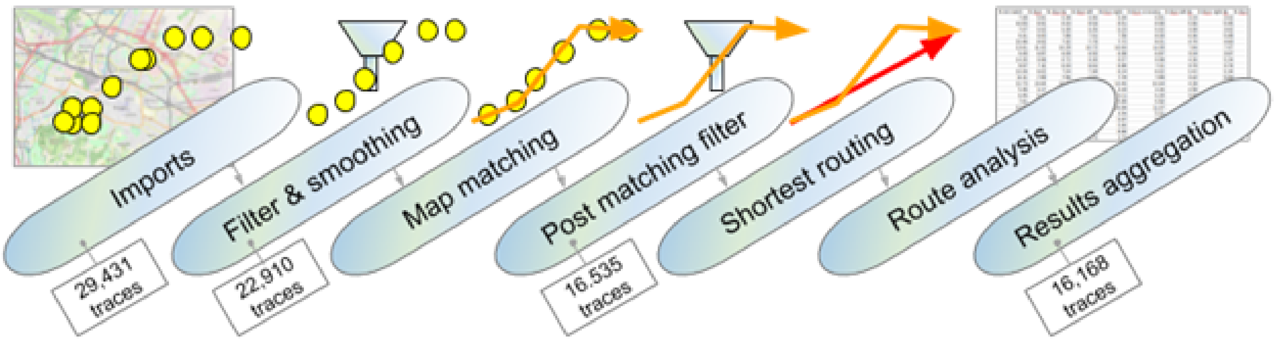


Fig.2 Evaluating cyclist attributes starting from GPS traces registered with Smartphones.

After a further filter that ensures the quality of the map-matching process, the shortest route connecting same link of origin and destination as the matched route has been identified for each trip. Subsequently, the matched routes have been analyzed - also from a dynamic point of view (see Rupi et al., 2020) - and compared with the respective shortest routes. There has been a further filtering of GPS traces that are suitable for the dynamic analysis. Finally, the attributes related to all the trips carried out from the same users have been aggregated in 5 different ways: mean, mean weighted by trips distances, median, standard deviation and mean absolute difference. These averaged attributes describe the experience of each user. After carrying out all the aforementioned pre-processing step the cyclist population is composed of 2,135 users (see tab.1), recording a total of 16,168 trips. This cyclist population belongs mainly to the firsts two age groups - 16-44 years - based on the phases of life described by Wittwer et al., 2014.

N users	N trips	Male	16→29	30→44	45→54	55→65	>65
2135	16168	46.50%	43.90%	34.90%	12.20%	8.30%	0.60%

Tab.1 Sample characterization in terms of size, gender and age - based on the phases of life described by Wittwer.

The following aggregate attributes have been obtained from the GPS trace analysis: trip length, trip duration and average speed; number of road priority changes per km; share of trip inside the center area, in roads reserved for cyclists, in mixed roads - shared with taxi, bus or pedestrians - and in low-priority roads; number of passed nodes, left turns, right turns and crossing - both in total and only in presence of a traffic light system (TLS) - per km; average numbers of maneuvers in the traveled intersections; all the route-attributes minus the same attributes referred to the shortest route; share of the shortest route length respect to the matched route; average in-motion speed; waiting time, as well as the registered waiting time minus the expected waiting time - based on the waiting times registered by the other users - at nodes, left turns, right turns, crossings - both in total and only in presence of a TLS - and at edges, per km. In addition to these attributes, the Bellamossa database contains additional information on the users such as: age, gender and number of recorded traces by each participant.

Tab.2 shows a statistical description of all the trip attributes aggregated for each cyclist as the median of the attributes related to their recorded trips. The table highlights also the dispersion of the various attributes referred to all cyclists. This suggests that cyclists have different habits and there might be the possibility of grouping the data set in some cyclist clusters. In particular, the average speed in motion is 4.85 ± 1.04 m/s, the average speed is 3.77 ± 1.11 m/s and the waiting time represents on average 14.6 % of the whole trip duration; these results are very similar to the dynamic results presented in (Rupi et al. 2020), where a different set of GPS-traces from Bologna, Italy, has been analyzed.

Attribute	Unit	Aver.	St.Dev.	Min	1stQu.	2ndQu.	3rdQu.	Max
Average length	m	2864	1607	189	1721	2505	3603	13650
Share of trip in the center area	%	42.57	35.51	0.00	0.00	40.92	71.62	100.00
Average speed	m/s	3.58	0.91	1.14	2.98	3.52	4.10	10.14
Number of priority changes	1/km	0.75	0.84	0.00	0.06	0.54	1.06	6.59
Share of reserved cycleway	%	23.79	20.10	0.00	6.98	20.07	36.47	100.00
Share of mixed road	%	30.52	18.43	0.00	16.84	27.91	41.27	100.00
Share of low-priority roads	%	69.91	22.70	0.00	54.37	72.59	89.08	100.00
Number of nodes	1/km	16.60	3.01	3.73	14.78	16.57	18.39	32.15
Number of left turns	1/km	2.11	0.98	0.00	1.45	2.02	2.68	6.97
Number of right turns	1/km	2.37	1.08	0.00	1.68	2.30	2.97	15.91
Number of crossings	1/km	11.36	2.81	0.00	9.45	11.25	13.18	28.37
Number of TLS nodes	1/km	2.82	1.60	0.00	1.70	2.76	3.73	10.92
Number of TLS left turns	1/km	0.39	0.45	0.00	0.00	0.31	0.60	4.14
Number of TLS right turns	1/km	0.38	0.38	0.00	0.00	0.34	0.60	3.07
Number of TLS crossings	1/km	1.89	1.21	0.00	0.98	1.82	2.68	8.32
Average number of maneuvers per node	/	10.30	1.19	6.00	9.64	10.22	10.82	22.25
Share shortest length	%	84.73	7.83	32.58	80.60	85.84	90.30	100.00
Number of priority changes*	1/km	-0.09	0.95	-6.18	-0.33	0.00	0.32	4.64
Share of reserved cycleway*	%	4.03	14.98	-59.80	-1.27	0.00	9.06	85.06
Share of mixed road*	%	-1.86	12.50	-72.05	-6.87	-0.31	4.16	52.28
Share of low-priority roads*	%	4.05	16.64	-63.12	-2.83	1.36	9.98	77.68
Number of nodes*	1/km	-0.73	2.33	-19.60	-1.75	-0.50	0.49	10.77
Number of left turns*	1/km	0.37	0.88	-5.74	-0.09	0.36	0.83	4.27
Number of right turns*	1/km	0.45	0.98	-5.17	-0.08	0.45	0.97	7.01
Number of crossings*	1/km	-1.59	2.33	-14.14	-2.70	-1.39	-0.27	10.11
Number of TLS nodes*	1/km	0.02	0.96	-5.13	-0.40	-0.04	0.43	7.50
Number of TLS left turns*	1/km	0.11	0.39	-2.73	-0.02	0.00	0.24	4.14
Number of TLS right turns*	1/km	0.11	0.36	-1.98	-0.01	0.00	0.28	2.66
Number of TLS crossings*	1/km	-0.22	0.81	-4.15	-0.57	-0.13	0.07	3.54
Average number of maneuvers per node*	/	0.08	0.87	-3.49	-0.31	0.04	0.39	5.64
Average in motion speed	m/s	4.70	0.92	1.54	4.13	4.63	5.19	11.11
Share of waiting time whole trip	%	14.71	9.96	0.00	7.56	13.13	19.88	71.74
Average waiting time per node	s	1.70	1.77	0.00	0.55	1.24	2.27	17.31
Average waiting time per left turn	s	1.94	4.67	0.00	0.00	0.00	1.88	67.00
Average waiting time per right turn	s	1.35	3.44	0.00	0.00	0.00	1.21	54.00
Average waiting time per crossing	s	1.61	2.11	0.00	0.33	1.00	2.15	35.73
Average waiting time per TLS node	s	4.12	6.32	0.00	0.00	2.25	5.70	96.30
Average waiting time per TLS left turn	s	3.22	8.95	0.00	0.00	0.00	1.00	85.00
Average waiting time per TLS right turn	s	2.82	9.29	0.00	0.00	0.00	0.00	122.00
Average waiting time per TLS crossing	s	4.44	8.20	0.00	0.00	1.71	6.00	147.17
Average waiting time per edge	s/km	13.17	25.75	0.00	0.69	5.34	14.34	334.25
Average waiting time per edge**	s/km	-	22.77	-67.66	-7.65	-4.01	0.92	316.39
Average waiting time per node**	s/km	-	30.62	-76.55	-13.16	-3.97	9.44	521.03

Average waiting time per left turn**	s/km	-	15.20	-57.42	-7.80	-2.59	3.49	242.91
Average waiting time per right turn**	s/km	-	8.94	-45.64	-2.73	-0.84	0.00	187.51
Average waiting time per crossing**	s/km	-	12.83	-34.14	-2.07	-0.71	0.00	472.56
Average waiting time per TLS node**	s/km	-	20.96	-65.45	-9.43	-3.28	4.01	239.87
Average waiting time per TLS left turn**	s/km	-	7.47	-25.23	-2.13	-0.33	0.00	187.51
Average waiting time per TLS right turn**	s/km	-	6.90	-32.55	-1.57	-0.44	0.00	115.30
Average waiting time per TLS crossing**	s/km	-	12.17	-49.67	-5.55	-1.94	1.54	213.77
Average waiting time whole trip**	s/km	-	39.48	-97.21	-16.49	-5.07	9.73	428.71
* Value referred to the matched trip minus the value referred to the shortest trip								
** real minus expected waiting time, based on the average waiting times of other cyclists that passed the same elements of the netwrork								

Tab.2 Descriptive statistics of cyclist attributes considered as the median of values referred to their performed trips.

From a first graphical analysis the Bologna cyclist data set does not clearly show spherical, symmetric, or convex clusters, therefore, a flexible distribution, like the CGHD, is preferred for the cyclist clusterization. The CGHD can capture differently shaped clusters, as well as, symmetric or spherical clusters.

4. Results

4.1 Data Analysis

Since the goal of the analysis is to find homogeneous groups of cyclists based on their cycling habits, the demographic variables have only been used for interpretation of the clusters after the analysis.

The total number of variables used for the analysis is 41 and the missing values have been imputed using multivariate imputation by chained equations (MICE) (Buuren and Groothuis-Oudshoorn 2010), i.e. each missing value has been imputed 5 times, obtaining 5 different complete data sets; the same analysis has been therefore performed on the 5 data sets, and the adjusted Rand Index (ARI) (Huber and Arabie 1985) has been used as a measure of pairwise cluster agreement in order to assess the robustness of the models, i.e. if all the partitions are similar, then the analysis is robust. The ARI is equal to 1 when there is perfect agreement between two partitions and the expected value is 0 for random classification: in the case study, the ARI was between 0.71 and 0.76, indicating a good level of robustness. The reported results refer to one of the data sets.

The analysis has been run using the software R (R Core team, 2020) by varying the number of clusters between 2 and 5, the BIC selected three clusters. The algorithm was initialized using a robust clustering technique called k-medoids (Kaufman and Rousseeuw 1990), and the package MixGHD (Tortora et al. 2019) has been used for the cluster analysis.

4.2 Cluster analysis results

The three clusters of cyclists present similar ages, but different trip attributes and cyclist choices and behaviors. The cyclist groups are composed of 806, 749 and 580 cyclists, with respectively 51%, 63% and 46% of women. According to the cluster's attributes, the three cyclist categories have been named as follow:

1. **Risky & Hasty** cyclists (RHC): they prefer going straight along the shortest path, traveling on unsafe roads and also large roads. This type of cyclist accepts also roads without reserved cycleway and a high density of intersections. The RHC is in average fast when in motion, but he/she loses time encountering many traffic lights, even if the average waiting time at intersections is fairly low, indicating that the RHC does often ignore the red signal.

- *Keywords: no deviations from the shortest path - low waiting times - fast*

2. **Inexperienced & Inefficient** cyclists (IIC): they travel at a low speed and they showed considerably higher waiting times respect to the other cyclists in all infrastructure elements (roads, intersections, traffic lights), probably because they are precautionary, but also afraid and not completely convinced to use a bike. They make deviations from the shortest path like the SIC group, but not to travel on safer roads with more exclusive cycleways, less traffic and fewer intersections.

- *Keywords: considerable waiting times - low speeds*

3. **Smart & Informed** cyclists (SIC): they know how to make deviations from the shortest path in an efficient way, searching frequently for roads with more reserved cycleways, fewer traffic and fewer intersections, thus gaining on both, safety and speed. They showed fairly low waiting times at intersections - like the RHC group - because they encounter less traffic lights and they probably pass also with the red signal (supposedly when it is safe).

- *Keywords: smart deviations from the shortest path for safer or faster roads - low waiting times - fast*

The radar graphs show the differences in terms of behavior (Fig.3) and trip characteristics (Fig.4) for the three different types of cyclists. The graphs contain the normal standardization of the first, second and third quartiles of the attributes referred to the cyclists of three clusters, - of which cyclist attributes are considered as the median value of the attributes referred to their recorded traces - in order to better visualize the differences in the same scale, while Tab.3 shows a descriptive analysis of the same attributes.

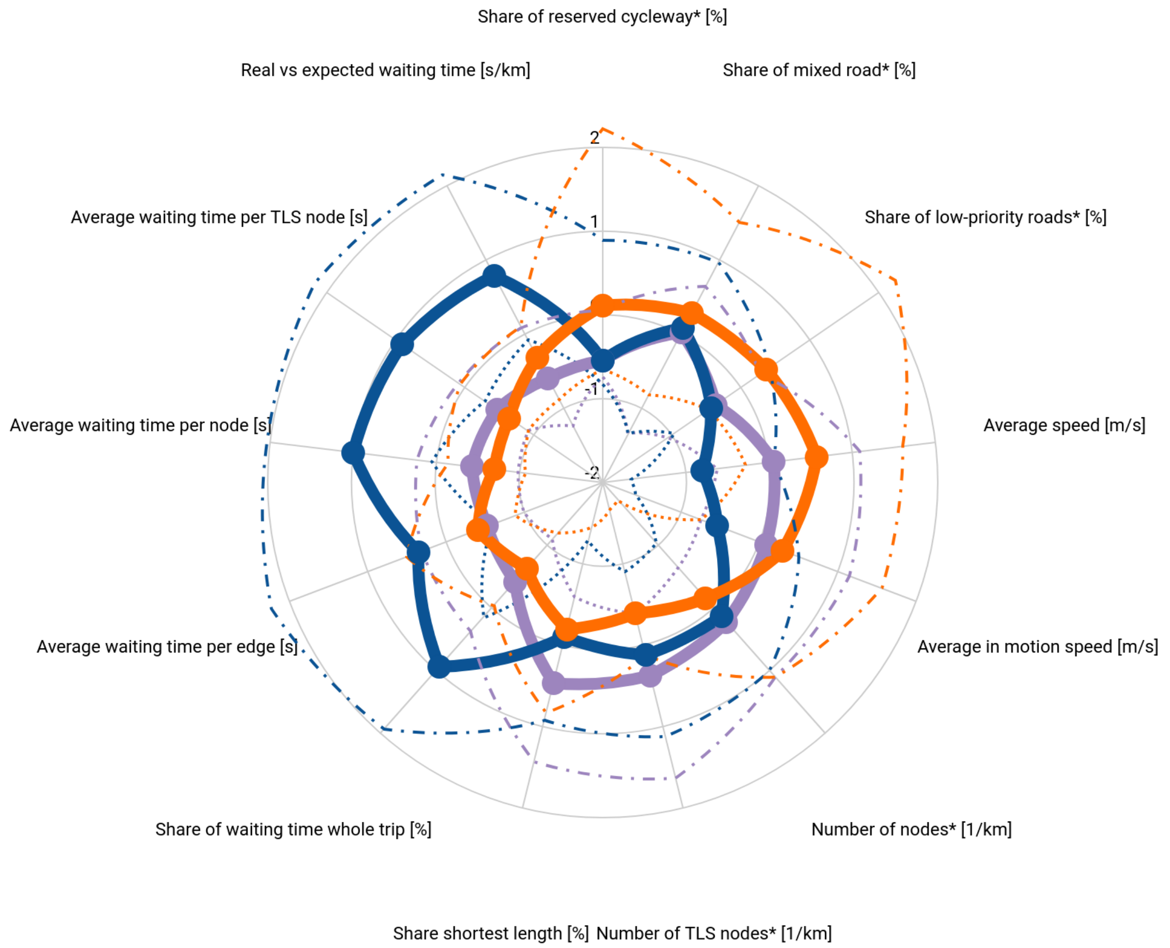


*Fig.3 Behavioral differences of the three cyclist types - RHC in purple, IIC in blue and SIC in orange. The dotted lines represent the first and third quartiles, while the solid line represent the median of cyclist attributes considered as the median of values referred to their performed trips. * Value referred to the matched trip minus the value referred to the shortest trip.*

The RHC type - who prefers to go straight on the shortest route - travel mainly on unsafe roads while encountering many intersections – of which most are with traffic lights – in particular in central areas of Bologna. The IIC type characterized by considerably high waiting times and low speed, indicating an insecure and cautious behavior. The other trips characteristics correspond to average values with respect the other cyclist types. The SIC type smartly deviates from the shortest route, often travels longer distances and changes frequently the priority of the road: his deviations contain safer roads with reserved cycleways, low priority roads and fewer intersections. The SIC also prefers to travel outside the center.

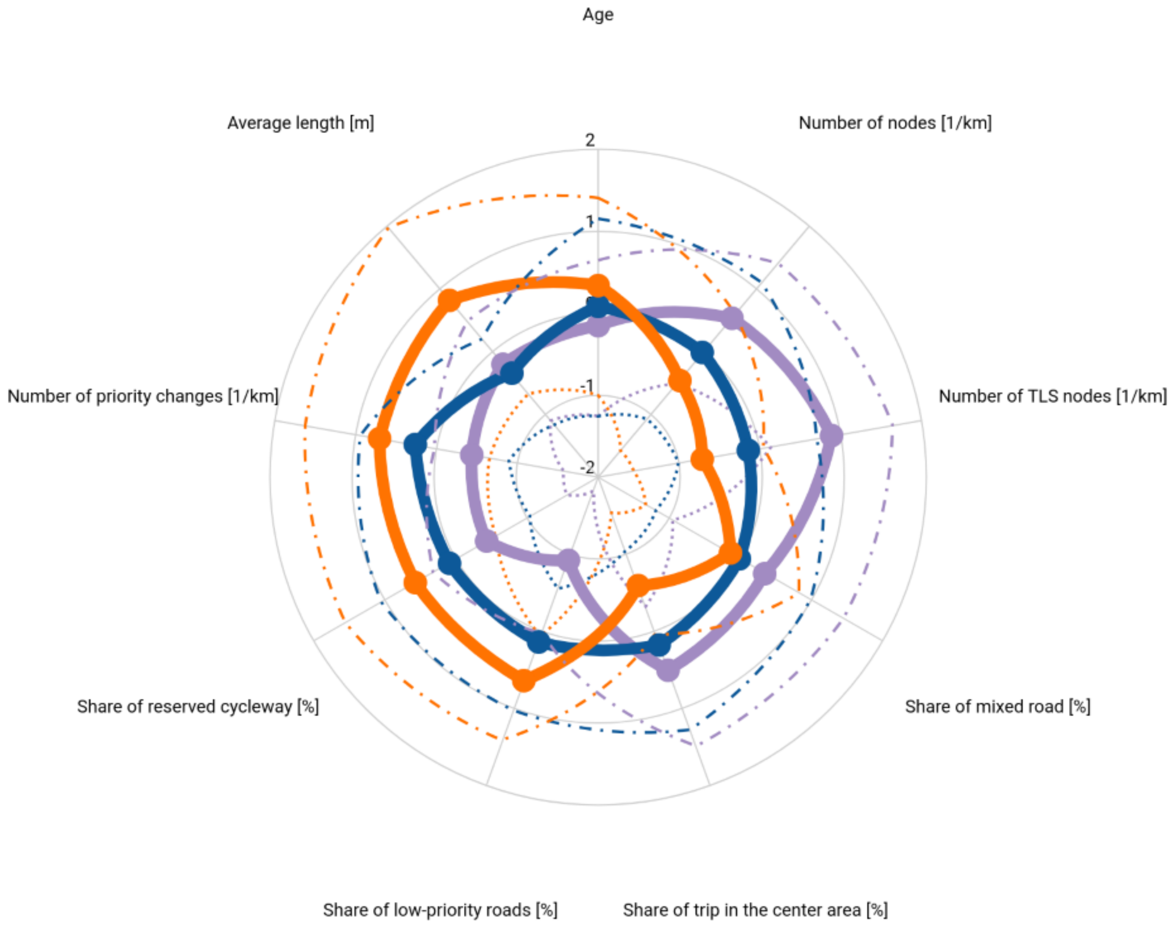


Fig.4 Differences in terms of trip characteristics of the three cyclist types - RHC in purple, IIC in blue and SIC in orange. The dotted lines represent the first and third quartiles, while the solid line represent the median of cyclist attributes considered as the median of values referred to their performed trips.

Attributes		**1st Quartile**			**2nd Quartile**			**3rd Quartile**		
Name	**Unit**	**RCU**	**IIC**	**SIC**	**RCU**	**IIC**	**SIC**	**RCU**	**IIC**	**SIC**
Share of reserved cycleway*	%	-1.2	-1.8	-0.7	0	0	4.1	3.8	9	17.3
Share of mixed road*	%	-7.9	-7.9	-5.2	-0.8	-0.5	0.6	2.5	4.2	7
Share of low-priority roads*	%	-4.9	-4.2	0	0.8	0.2	6.7	6	7.2	22
Average speed	m/s	3.2	2.6	3.4	3.6	3.1	3.9	4.2	3.6	4.5

Average in motion speed	m/s	4.3	3.9	4.3	4.7	4.4	4.8	5.2	4.9	5.4
Number of nodes*	1/km	-1.3	-1.7	-2.4	-0.4	-0.5	-0.8	0.5	0.4	0.5
Number of TLS nodes*	1/km	-0.2	-0.4	-0.7	0.1	0	-0.2	0.6	0.4	0
Share shortest length	%	82.8	79.7	79	87.4	84.9	84.5	91.7	89.4	89
Share of waiting time whole trip	%	5.8	14.6	5.1	10.6	20.2	9.1	16.2	27.2	13.3
Average waiting time per edge	s	0	2.9	0.8	3.4	10	4.3	9.6	23.9	11.1
Average waiting time per node	s	0.4	1.5	0.3	1	2.5	0.7	1.7	3.6	1.3
Average waiting time per TLS node	s	0	2.1	0	1.5	6.1	0.9	3.4	10.3	3.3
Real vs expected waiting time	s/km	-22.9	-3.7	-17	-12.1	10.8	-7.7	-0.8	33.3	-0.6
Age	-	13	15	16	30	31	35	39	43	45
Number of nodes	1/km	15.5	14.7	13.8	17.2	16.5	15.7	18.9	18.4	17.5
Number of TLS nodes	1/km	2.8	1.6	0.9	3.5	2.5	1.8	4.5	3.4	2.8
Share of mixed road	%	19.2	16.5	15	31.9	26.1	26	45.6	39.4	38.5
Share of trip in the center area	%	31	0	0	56.3	42.6	13.3	82.6	77.8	41.5
Share of low-priority roads	%	41.3	61.3	72.5	56.7	77.4	85.2	72.7	91	95.4
Share of reserved cycleway	%	2.9	8.1	15.5	12.3	21	28.9	28.5	37.3	43.6
Number of priority changes	1/km	0	0.2	0.4	0.3	0.6	0.8	0.7	1.1	1.4
Average length	m	1680	1612	2075	2346	2286	3325	3318	3038	4990
* Value referred to the matched trip minus the value referred to the shortest trip										

Tab.3 Descriptive analysis of the clusters with cyclist attributes considered as the median of values referred to their performed trips.

5. Conclusions

The present study identified three types of cyclists by applying a cluster analysis to attributes calculated from GPS traces. The data set is composed of 16,168 GPS traces recorded between 7-10 am on work days from April to September 2017. Based on the results of the cluster analysis, the types have been named **Risky & Hasty** cyclists (RHC), **Inexperienced & Inefficient** cyclists (IIC) and **Smart & Informed** cyclists (SIC). The novelty of the present study, is the use of crowdsourced GPS data that allows to analyze a large data sample as objective, revealed survey data. The pre-processing of the GPS traces has produced a large and variegate set of attributes which characterize the cyclist experience, leading to a meticulous description of the different types of cyclists. The most significant attributes in the cyclist clusterization are: the amount of trip deviation made for improving the cyclist experience in terms of both safety and speed as well as the waiting times detected in various parts of the road network.

Rather than providing threshold values to split other cyclist database into the three groups, a model-based approach has been used. Each cluster is modeled using a CGHD distribution with different parameters. The model can be used to find the belonging probability to each cluster of cyclists in the same area not included in the original sample.

The proposed paper presents on what these groups differ and how much, thus presenting the differences in terms of experience, performances, choices and behavior of each group of cyclists for design both new surveys and infrastructures in the city - adapted to each group of cyclists - as well as initiatives that can allow to improve the trip experience of certain cyclists, making it safer and more efficient as well as reducing barriers for cycling.

One current limitation of the data set is that it may not be representative. In future research it is possible to attribute more weight to the data of underrepresented parts of the population.

Acknowledgements

We are grateful to SRM Bologna srl for providing the GPS data of the Bella Mossa Campaign 2017.

References

Betterpoints. (2020). https://www.betterpoints.ltd/behaviour-change/ Accessed 1 May 2020.

Bernardi, S., La Paix-Puello, L., Geurs, K., (2018). Modelling route choice of Dutch cyclists using smartphone data. *J. Trans. Land Use 11* (1), 883–900.

Broach, J., Dill, J., Gliebe, J., (2012). Where do cyclists ride? A path choice model developed with revealed preference GPS data. *Transp. Res. Part A* 46, 1730–1740. 10.1016/j.tra.2012.07.005

Browne, R. P., McNicholas, P. D. (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics*, 43(2), 176-198. 10.1002/cjs.11246

Buuren, S. V., Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1-68. 10.18637/jss.v045.i03

Cabral, L, Kim, A. M., (2020). An empirical reappraisal of the four types of cyclists. *Transportation Research Part A: Policy and Practice*, Volume 137, 206-221. 10.1016/j.tra.2020.05.006

Charlton B., Sall E., Schwartz M., Hood J. (2011). Bicycle Route Choice Data Collection using GPS-Enabled Smartphones, *TRB 2011 Annu. Meet.* 1– 10.

Chen P., Shen Q., Childress S., (2018). A GPS data-based analysis of built environment influences on bicyclist route preferences, *Int. J. Sustain. Transp.* 12(3), 218–231. 10.1080/15568318.2017.1349222

Clarry, A., Faghih Imani, A., Miller, E.J., (2019). Where we ride faster? Examining cycling speed using smartphone GPS data. *Sustainable Cities and Society 49*, 101594. 10.1016/j.scs.2019.101594.

Damant-Sirois G., Grimsrud M., El-Geneidy A. M. (2014). What's your type: a multidimensional cyclist typology. *Transportation*. 41(6), 1153–1169. 10.1007/s11116-014-9523-8

Dill, J., (2009). Bicycling for Transportation and Health: The Role of Infrastructure. *J. Public Health Policy* 30, 95–110. 10.1057/jphp.2008.56

Dill J., McNeil N. (2013). Four Types of Cyclists? Examination of Typology for Better Understanding of Bicycling Behavior and Potential. *Transportation Research Record: Journal of the Transportation Research Board*. 2387, 129-138. 10.3141/2387-15

Dill J., McNeil N. (2016). Revisiting the Four Types of Cyclists: Findings from a National Survey. Transportation Research Record: Journal of *the Transportation Research Board*. 2587, 90-99. 10.3141/2587-11

Eclipse SUMO. (2020). http://sumo.sourceforge.net/userdoc/ Accessed 1 May 2020.

Everitt, Brian S., et al. (2011). Cluster analysis. *John Wiley & Sons.*

Flügel S., Hulleberg N., Fyhri A., Weber C., Ævarsson G. (2017). Empirical speed models for cycling in the Oslo road network. *Transportation (Amst)*. 1.

Francke A., Anke J., Lißner S., Schaefer L.-M., Becker T., Petzoldt T. (2020). Are you an ambitious cyclist? Results of the cyclist profile questionnaire in Germany. *Traffic Inj. Prev.* 0(0), 1–6.

Gatersleben B., Haddad H. (2010). Who is the typical bicyclist? *Transp Res Part F Traffic Psychol Behav.* 13(1), 41–48. 10.1016/j.trf.2009. 10.003

Geller, R. (2006) Four Types of Cyclists. Portland Bureau of Transportation, Portland, Ore. http://www.portlandoregon.gov/transportation/article/264746 Accessed 31 May 2020.

Griffin, G.P., Jiao, J., (2015). Where does bicycling for health happen? Analyzing volunteered geographic information through place and plexus. *J. Transp. Health 2*, 238–247. https://doi.org/10.1016/j.jth.2014.12.001

Hood, J., Sall, E., Charlton, B., (2011). A GPS-based bicycle route choice model for San Francisco. California. *Transp. Lett. Int. J. Transp. Res.* 3, 63–75. https://doi.org/10.3328/TL.2011.03.01.63-75

Hubert, L., Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1), 193-218. 10.1007/BF01908075

Joo S. Oh C. (2013). A novel method to monitor bicycling environments. *Transp. Res. Part A Policy Pract.*, 54, 1–13.

Kaufman, L., Rousseeuw, P. J. (1990). Partitioning around medoids (program pam). *Finding groups in data: an introduction to cluster analysis*, 344, 68-125.

Kroesen M., Handy S. (2014). The relation between bicycle commuting and non-work cycling: results from a mobility panel. *Transportation*. 41(3), 507–527. 10.1007/s11116-013-9491-4

Laranjeiro, P.F., Merchán, D., Godoy, L.A., Giannotti, M, Yoshizaki, H.T.Y., Winkenbach, M., Cunha, C.B., (2019). Using GPS data to explore speed patterns and temporal fluctuations in urban logistics: The case of São Paulo, Brazil. *Journal of Transport Geography*, 76, 114-129, 10.1016/j.jtrangeo.2019.03.003

Lu, W., Scott D., M., Dalumpines, R. (2018). Understanding bike share cyclist route choice using GPS data: Comparing dominant routes and shortest paths. *Journal of Transport Geography*.71, 172-181. 10.1016/j.jtrangeo.2018.07.012

Manum, B., Arnensen, P., Nordström, T., Gil, J. (2018). Improving GIS-based models for bicycling speed estimations. *Eur. Transp. Conf.*

McLachlan G. J., Basford K. E. (1988). Mixture models: inference and applications to clustering. *Statistics, textbooks and monographs. 84.*

McNicholas, P.D. (2016). Model-Based. *Journal of Classification. 33, 331-373. 10.1007/s00357-016-9211-9*

Menghini, G., Carrasco, N., Schüssler, N., Axhausen, K.W., (2010). Route choice of cyclists in Zurich. *Transp. Res. Part A* 44, 754–765.

OpenStreetMap. (2020). https://www.openstreetmap.org Accessed 1 May 2020.

Pritchard R. (2018) ‘Revealed preference methods for studying bicycle route choice—a systematic review’, *Int. J. Environ. Res. Public Health*, 15, 3.

Pritchard, R., Frøyen, Y.K., Bernhard, S., (2019). Bicycle Level of Service for Route Choice—A GIS Evaluation of Four Existing Indicators with Empirical Data. *Int. J. Geo-Inf.* 8, 214. doi:10.3390/ijgi8050214

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. htt10.3390/ijerph1503047010.3390/ijerph15030470p://www.R-project.org/. Accessed 31 May 2020.

Rupi, F., Poliziani, C. Schweizer, J. (2019). Data-driven bicycle network analysis based on traditional counting methods and GPS traces from smartphone. *ISPRS International Journal of Geo-Information*, 8. 10.3390/ijgi8080322

Rupi, F.; Schweizer, J. (2018), Evaluating cyclist patterns using GPS data from smartphones. IET *Intell. Trans. Syst.*, 12, 279–285. 10.1049/iet-its.2017.0285

Rupi, F., Schweizer, J., Poliziani, C., (2020), Analysing the dynamic performances of a bicycle network with a temporal analysis of GPS traces. *Case Studies on Transport Policy.* 10.1016/j.cstp.2020.05.007

Schweizer J., Sumopy. (2020). https://sumo.dlr.de/docs/Contributed/SUMOPy.html Accessed 1 May 2020.

Schweizer, J., Bernardi, S., Rupi, F., (2016), Map-matching algorithm applied to bicycle global positioning system traces in Bologna, *IET Intell. Transp. Syst.*, 10, 244–250. 10.1049/iet-its.2015.0135

Schweizer, J, Rupi, F, Poliziani, C. (2020). Estimation of link-cost function for cyclists based on stochastic optimization and GPS traces. *ITE Intelligent Transport Systems*, in press.

Strauss, J., Miranda-Moreno, L.F., (2017). Speed, travel time and delay for intersections and road segments in the Montreal network using cyclist Smartphone GPS data. *Transp. Res. Part D: Trans. Environ.* 57, 155–171. doi:10.1016/j.trd.2017.09.001.

Strauss J., Zangenehpour S., Miranda-Moreno L. F., Saunier N. (2017). Cyclist deceleration rate as surrogate safety measure in Montreal using smartphone GPS data, *Accid. Anal. Prev.*, 99, 287–296. 10.1016/j.aap.2016.11.019

Tortora C., ElSherbiny A., Browne R.P., Franczak B.C., McNicholas P.D. (2019). MixGHD: Model Based Clustering, *Classification and Discriminant Analysis Using the Mixture of Generalized Hyperbolic Distributions.* https://CRAN.R-project.org/package=MixGHD, R package version 2.3.1. Accessed 31 May 2020.

Tortora, C., Franczak, B. C., Browne, R. P., McNicholas, P. D. (2019). A mixture of coalesced generalized hyperbolic distributions. *Journal of Classification*, 36(1), 26-57. 10.1007/s00357-019-09319-3

Watkins K. E. and LeDantec C. (2016). Using Crowdsourcing to Prioritize Bicycle Route Network Improvements. *Report FHWA-GA-16-1439'*

Wittwer, R., Zwangsmobilität, Verkehrsmittel. (2014). Orientierung junger Erwachsener: Eine Typologisierung, Schriftenr. Dresden: Technische Universität Dresden Fakultät Verkehrswissenschaften „Friedrich List" Institut für Verkehrsplanung und Straßenverkehr Geschäftsführender Institutsdirektor: Univ.-Prof. Dr.-Ing. Christian Lippold.

Zimmermann, M., Mai, T., Frejinger, E., (2017). Bike route choice modelling using GPS data without choice sets of paths. *Transp. Res. Part C* 75, 183–196.