Vedernikov *et al. Open Geospatial Data, Software and Standards* (2016) 1:12
DOI 10.1186/s40965-016-0012-1

Open Geospatial Data, Software and Standards

ORIGINAL ARTICLE Open Access

The Hitchhiker's guide to the pick-up locations

Oleksii Vedernikov*, Lars Kulik and Kotagiri Ramamohanarao

Abstract

Background: Hitchhiking is a well-known form of transportation, which has been extensively studied from a sociological perspective. While this tourism approach has been popular in some countries for decades, the question of making hitchhiking fast has never been studied from a comprehensive point of view. In this research, we aim to find the best locations for hitchhiking, which may be used in travel recommender systems.

Methods: For this purpose, we study the relationships between certain spatial properties and waiting time at pick-up locations, estimate statistical significance of several factors according to various countries.

Results: We extracted features for the data from hitchhiking logs, analyzed their importance and built corresponding classification and regression models for estimating hitchhiking experience with a sufficient accuracy.

Conclusions: This paper is the first to analyze important spatial factors for a successful hitchhiking trip and construct models to predict the quality and waiting time at hitchhiking locations.

Background

Motivation

Hitchhiking is the first well-known form of ridesharing that became popular in 20th century in the US, and wide spread all around the world. It is an act of soliciting and getting rides from random drivers passing by the road without prior arrangement. For example, an English Wikipedia article about hitchhiking [1] contains references to 44 movies, 21 books, 34 songs, 19 notable hitchhikers, which shows the significance of this cultural phenomenon. Besides, hitchhikers constitute a large international community. For example, a website *Hitchlog* contains information about 1,272 hitchhikers' trips with a total length of more than 2,300,000 km all over the world.

As a goodwill exchange, hitchhiking promotes values for sharing cost and helping unknown people. It facilitates friendship among various people from different cultural backgrounds. Since hitchhiking involves uncertain and spontaneous travels, it is usually considered an adventurous way of travelling. In addition, it can be done on any road with traffic. Even though personal safety is a common concern about this type of travel, surveys show that only 1% of people who hitchhiked report negative experience of hitchhiking [2].

*Correspondence: ovedernikov@student.unimelb.edu.au
University of Melbourne, Melbourne, Australia

Hitchhiking has attracted substantial number of research papers since 1960s, and most of these studies are concerned with its social and cultural aspects. Despite its prominence and importance, to the best of our knowledge, the process of hitchhiking has never been analysed from a computational point of view. Our objectives, namely making hitchhiking fast and efficient, have never been studied.

Research problem

The process of hitchhiking includes different stages. First, having in mind a final location, a hitchhiker chooses a route and decides where to start seeking a ride. When a car stops, both the driver and the hitchhiker negotiate their travel intentions. If reaching the hitchiker's destination requires several rides, the same process is repeated, and the number of these steps may vary from one to many due to the uncertain nature of hitchhiking.

Even though there are several factors for a successful hitchhiking ride, the location is considered the most important factor for a successful ride [3]. Certain location properties make it easier for drivers to stop and have enough traffic to increase the probability of a successful ride for a hitchhiker. A good location is determined by two main factors: a short average waiting time for catching a ride to the desired destination, and a small probability of an unusually long ride. Therefore, the choice of a

hitchhiking location is crucial for a successful hitchhiking trip. This paper is the first one to identify these factors by analysing hitchhiking logs.

The main research question of this paper is:

- Which factors influence the popularity of some locations over others, and subsequently lead to a positive travel experience?

To answer this question, we investigate the hitchhikers' website *Hitchwiki* that provides thousands of recorded experiences with location ratings, waiting time, and comments. Using *OpenStreetMap*, we investigate which types of roads are most frequently used by hitchhikers. We also investigate the waiting time and rating of a given location based on several spatial factors taken from hitchhikers' descriptions.

The key contributions of this paper are:

- More than 9000 locations were analysed according to 11 types of roads and 5 different facilities. The most prominent features are gas stations, traffic lights and bus stops, which usually attract more hitchhikers. Among the different types of roads, ramps on motorways are extremely popular, while minor types of regional roads are rare.
- We performed a text analysis of the user descriptions and comments, and found dependencies between road types and certain facilities on them. For example, gas stations are popular hitchhiking points on motorways, while bus stops may be significant pick-up points on regional roads.
- We have developed a feature set for classifying good and bad ratings of locations based on location features as well as a regression model to predict exact ratings.

The results of this paper describe certain properties needed for a successful hitchhiking location, show how hitchhiking experience is different for various roads, and how users' comments can improve the accuracy of a recommender system. These results are highly important to assist hitchhikers in helping them to plan an optimal trip and can assist in developing a trip planning application. In addition, infrastructure planners and local authorities, can consider these results to improve hitchhiking facilities, for example making special signs with stops or providing supplementary amenities for hitchhikers.

The rest of the paper is organized as following: "Literature review" section contains aliterature review on hitchhiking and other forms of ridesharing. "Methodology" section gives a description of whole hitchhiking process and "Results and discussion" section contains a description of the Hitchwiki dataset and shows the results of experiments. Finally, "Conclusions" section provides conclusions and future directions of hitchhiking-related researches.

Literature review

Hitchhiking

As a social and cultural phenomena, hitchhiking became popular in 1940s and 1960s and consequently attracted many studies from sociological and psychological point of view [4]. For example, [5] experimentally showed that females attract approximately 3 times more cars to stop than males, while eye contact doubled this chance for both genders. This work also proves importance of quite specific variables, for example beards, thumbing up for males and bust size accentuation for females. Despite quite common stereotypes of unsafety, a 2014 survey in the Netherlands [2] states that 84% of hitchhikers have positive or predominantly positive experience, contrary to the facts that people commuting by public transport are usually unhappy. One of the reasons is a recent scientific exploration that making contacts with strangers can make people happy. Among reasons which lead drivers to pick up passengers, the most prominent are desire to help them (61.3%), having experience as hitchhikers in the past (49.5%) and opportunity to get connected with new people (45.0%).

Recently, a few sociological studies were conducted to understand links between drivers' willingness and factors like: hair colour [6], bust size [7], make up [8], humor [9], weather conditions [10], and even various types of hijab [11]. Usually, these studies make hypotheses, and set hitchhiking experiments in different conditions according to hypotheses. For example, [11] prove that liberal dress led 5 times more cars to stop than a conservative one, and [10] state that sunny days facilitate drivers to stop.

Some of the hitchhikers provide comprehensive details of their travels. For example, Frank [12] covered 132620 km by hitchhikikng in 2008 rides with average of 66 km per ride. Average waiting time for a ride is 19 min, or 17.3 s per kilometre.

In some areas hitchhiking has been recently promoted by local communities and organizations, there are some modern studies about hitchhiking. A social project Hopista (previously known as Lawrence OnBoard and CarmaHop), which has a goal to promote hitchhiking as the basis of a community ridesharing service through social initiatives and special smartphone application to match passengers and drivers. Their experiments in Northern Kansas, US showed that the median time was 6 mins, and nearly 95% got a lift within 30 mins. Another example is NederlandLift [2] project, which aims to promote hitchhiking lifts due to benefits in financial aspect, climate change, congestion and parking problems. Additionally, they claim that hitchhiking promotes social values as opposition to isolationism, making travels more

fun, spontaneous and sociable. The similar organization has been started in Austria [13]. These works, supported by local governments, prove that even though hitchhiking may have some negative connotation in the past, it has wide opportunities for development and to solve current transportation problems. However, they all have lack of promotion, and one of ways to increase their popularity is to develop a hitchhiker assistant application.

To summarize, hitchhiking is considered as a compound of travelling, ridesharing, adventure and cultural exchange, which makes it interesting to study. It is shown that success of hitchhiking is related to many factors, including social, gender, cultural aspects, which are widely studied earlier. However, spatial and computational aspects of hitchhiking have not been studied before to the best of our knowledge, and this paper aims to fill this gap. In the "Methodology" section, a few descriptions of good hitchhiking locations will be presented.

Other forms of shared transportation

Hitchhiking may be characterised as a special type of ridesharing. In this subsection, we will review related works about other forms of ridesharing. Properties of different forms of ridesharing are summarized in Table 1.

Furuhata et al. [14] provide a comprehensive overview of ridesharing systems. 39 matching agencies were divided into 6 classes according to two main criteria: *primary search criteria* and *target market*. The former is related to the data needed to match drivers and passengers by a system, while the latter is divided into 3 categories: on-demand, commute and long-distance. Among other classes, the *flexible carpooling* has the most similarities with hitchhikikng, due to no prearrangement and coordination on the spot. Pricing could be catalog-based (the price is determined by the participants while listing), rule-based (usually proportional to the distance plus initial rate), and negotiation-based (determined by users, not by the system). Historically hitchhiking is associated with free ride, but current carpooling schemes usually involve fair division of travel costs between all participants.

A regional-spread form of ridesharing is slugging or casual carpooling, which became popular in four USA areas due to restrictions on car capacity at High-occupancy vehicle (HOV) lanes, which made more drivers and passengers show interest in ridesharing [15]. Specific spots on the way to tolled highways have become popular locations for picking up passengers to commute. Ma and Wolfson [16] prove that finding the optimal slugging plan to minimize total distance is NP-complete. In addition, the authors propose a few heuristics and show their effectiveness in terms of travel time. However, this work does not aim to find the optimal locations to form slugging pools, therefore our research can also contribute to this area by suggesting the best locations for slugging. Overall, slugging is the most similar form to hitchhiking. Like public transport, it assumes no prior arrangement between a driver and a passenger, but public transport has much more complicated demand modelling [17]. However, the difference with hitchhiking lies in regular commuting character, and predefined locations of start and end points. Investigation of convenient hitchhiking locations would benefit a potential spread of slugging with the help of finding new locations.

Therefore, carpooling in its different forms includes the problem of finding the best location to pick up passengers. Whereas a few slugging areas solve this problem by using prearranged locations as parkings near HOV lanes, general solution of finding the best locations has not been addressed yet.

Results summarized in Table 1 imply that hitchhiking has no prior arrangement between a passenger and a driver and does not involve change of an initial route, which is similar to both public transport and slugging. Hitchhiking usually involves one-way trips, like taxi services, while other forms of ridesharing allow an option to arrange the return trip too. The two main differences with all other ridesharing types include mostly free rides and intraurban character of travels.

Even though there is a substantial number of works on social analysis of hitchhiking, and different aspects of various forms of shared transportation, to the best of our knowledge, there is not a single work which researches the best locations for hitchhiking from a comprehensive point of view, and we address it in this paper.

Methodology

General problem description

The whole process of hitchhiking is described below. A hitchhiker has an initial location, a starting point, and has

Table 1 Properties of different ridesharing types

Ridesharing type	Prior arrangement	Recurring trip	Route change	Cost sharing	Urban
Public transport	NO	YES	NO	YES	YES/NO
Taxi ridesharing	YES	NO	YES	YES	YES
Carpooling	YES	YES/NO	YES	YES	YES
Slugging	NO	YES/NO	NO	YES	YES
Hitchhiking	**NO**	**NO**	**NO**	**NO/YES**	**NO**

a goal to go to a destination location, usually another city. Overall, the strategy is to hitchhike optimally in terms of minimizing one or more of the following:

- travel time (sometimes to a few hours)
- reliability (one route may have a constant number of cars at any time, while another may have high and low peaks)
- danger (experience of talking to strangers may be unpredictable)
- discomfort (for example, a stop with shelter, fresh water and fast-food nearby is better than an isolated location on a country road)
- cost (in some countries hitchhiking may involve payment)

Cost is more a matter of personal negotiation or cultural traditions: in Germany hitchhiking is usually free of charge, while in Fiji it is more common to pay a standard bus fare to a driver. Popular belief of insecurity is the main reason why there are not many hitchhikers, which contradicts to the actual data about dangers or risks while hitchhiking [2, 18]. Overall, risks and cost are dependent mostly on regional specific factors, so we may assume that they uniformly distributed if we narrow down the research area to certain countries. Comfort means that a hitchhiker should have basic facilities like fresh water or shelters, since waiting and travelling time could be long. Discomfort to a certain extent is not be a serious issue since a hitchhiker is already prepared to a travel and is aware of it, and a hitchhiker is able not to accept an offer or to continue travel by public transport.

We focus on the most important parameters, namely waiting time and its reliability, and aim to minimize them. In this research, we narrow down the area to the waiting time at pick-up locations, which is the most distinctive feature of hitchhiking comparing to other modes of transportation.

Considering a hitchhiker with an initial position and destination point, the total travel time consists of the following components:

1. the walking time from an initial location
2. the waiting time at a location (*)
3. the travel time to the next location (*)

Steps *2–3* are repeated until the hitchhiker reaches their destination. We will call this process as *hitch cycle* with *waiting* and *riding* components, and the stops in *hitch cycle* call *transitional* stops. Since there is unpredictable waiting time at all locations, it makes estimating travelling by hitchhiking different from a usual routing problem. Amount of such iterations may vary a lot, from 1 to many per hitchhiking day, and is dependent on many factors: how optimal is the chosen route in terms of number and destination of cars, how far they will go, how fast they could drive on certain roads etc. Even if cars' speed is high, but they do not go too far, a few extra stops may require substantial additional time to the travel, making it not-efficient. A hitchhiker usually cannot ask a driver to drive further than his destination or change his route, he can just ask him to stop earlier.

Hitchhiking location properties

A successful hitchhike is influenced by many parameters. First, there are factors which are related to person's gender, appearance, presence of sign etc. The success is also dependent on cultural values, spread of hitchhiking in a particular region. These points were mentioned in the "Literature review" section, but the most important factor is a location itself [3]. The quality of a hitchhiking location may be looked from various points of view, for example:

- How easy it is for cars to stop. In the middle of a motorway, there is no chance that cars could stop right after noticing a hitchhiker.
- How likely that cars are going in the direction which is useful for the hitchhiker. Most of the hitchhiker do long rides, so points with most of local traffic are senseless.
- Reachability of a point by public transport for locations to start hitchhiking. Usually hitchhikers start travelling in cities, and they need to reach highways.
- Location equipped with some basic facilities like water or shelter. Even though waiting time in a proper location should not be long, these amenities may be crucial for certain groups of travellers.
- Location prominence for drivers. For example, certain countries like the Netherlands have special stops with corresponding signs for hitchhikers. Therefore, information about these specific locations may be well-known among drivers.

Also, a several Internet-resourses ([3, 19–21]) contain information about characteristics of the best locations to hitchhike, namely good visibility of a passenger, cars' speed is not high, an optimal traffic, and a safe, easy, legal and obvious place to pull over. For example, good locations are located near gas stations, commercial rest areas, public highway on ramps, on-the-road restaurants, parking areas, land borders etc.

However, these observations have not been proved on a large scale. Our goal is to validate these factors, and find out how they influence the waiting time and rating of a location based on information in hitchhiking logs. For this reason, we need to analyse how close locations are to the different facilities, and how a distance to each

of the road facilities influences resulting performance in ratings and waiting time. We need to find the correlation between waiting time and rating, and see how it is influenced by the credibility of the assessment. We also need to find relationships between users' comments and location attributes. In addition, since hitchhiking character, traditions and habits vary a lot, a comprehensive study on different datasets should be done.

Since the main goal is to build a recommender system for hitchhikers, which is to be designed, we propose two possible solutions. One is a classification of location into two classes: those which may be used by hitchhikers, and those to be avoided, and regression model to predict the rating of a location. All the features selected and analysed on previous step should be used in both.

Results and discussion

Dataset description

We use Hitchwiki maps [19] dataset. There are 21,562 rated points (on 12 August 15), and they are not uniformly covered across all locations in the world. For example, whole Africa contains only 167 points, the top-10 countries sorted by the number of points are provided in Table 2.

These results conform with the difference in the spread of hitchhiking culture among other countries as well as popularity of particular *Hitchwiki* website. To make an extensive analysis, we do all the experiments described below with top-6 countries: Germany, France, Poland, Netherlands, United Kingdom and United States. In this section, we show a graph for a single country due to space constraints and provide the graphs for remaining countries in the Additional file 1.

Since the dataset is related to Volunteered Geographic Information (VGI) crowdsourced by users, its credibility should be considered. An extensive recent survey about quality assessment of VGI [22] summarizes all recent findings in the field. VGI data has been succesfully applied in many areas like discovering Points of Interest (POI) for estimating urban land use for urban planners [23] or travel reccomendations [24]. Even though there are 2525 points from Germany in the dataset, they have various levels of credence. Some locations may have only one vote, while others may be assessed 20–30 times and thus have higher confidence. For example, among 2525 spots from Germany, only 107 have 10 or more rating votes. This parameter of minimum amount of votes of a location would be defined as N_{min}, and will be assessed later.

Table 2 Top-10 countries from *Hitchwiki* dataset

Country	No. of spots
Germany	2525
France	2379
Poland	1282
Netherlands	997
United Kingdom	888
United States	793
Spain	675
Czech Republic	634
Russia	591
Romania	530

Users are able to create new points and edit existing points, assessing their rating and waiting time, and add comments about them. Since there are no regulations about assessing locations, users may vote for both "hitchability" rating of a location (on the scale from 1 to 5, from "Senseless" to "Very good") and its waiting time (5, 10, 15 etc. minutes). Since a point may be assessed a few times, the integer ratings become continuously distributed in the interval $[1..5]$, and waiting times - in the half-interval $[5..\infty)$.

Some of the locations have many detailed comments, and usually they comprise descriptions of some aspects such as: how good it is; which directions from this spot most of cars go; how long it takes to catch a car there etc. Consequently, text analysis of comments would be another future direction with a potential improving of the accuracy of the hitchhiking recommender system, and some initial results are presented below.

Users can vote spots' rating and/or write waiting time of their experience. Due to users preferences, reason, there are roughly two times more ratings than waiting time in the dataset. Figure 1a and 1b show that there is a moderate correlation between them for different N_{min}.

Figure 1a shows that there are not enough points with higher N_{min} values for some countries, and generally the correlation between waiting time and rating is moderate.

Figure 1b shows many points with integer ratings for $N_{min} = 1$ due to the nature of the data, which allow users to rate locations only with integers from 1 to 5. Next figures exhibit much less points along integer lines because average values of longer arrays of integers are less likely to be integers, and another observation is significant decrease of points with ratings lower than 2. This tendency is also illustrated in the Fig. 2: the locations with higher ratings are more likely to attract new people, while bad ratings, especially with an explicit description, may be a red flag for other users. Figure 2 illustrates the tendency of users to use more often locations which have already got high ratings. Interestingly, Fig. 3 shows no significant relationships between waiting time and ratings. The reason may be related to general uncertainty of hitchhiking: even at a good locations sometimes a hitchhiker needs to wait long, so the average waiting time is not decreasing significantly.

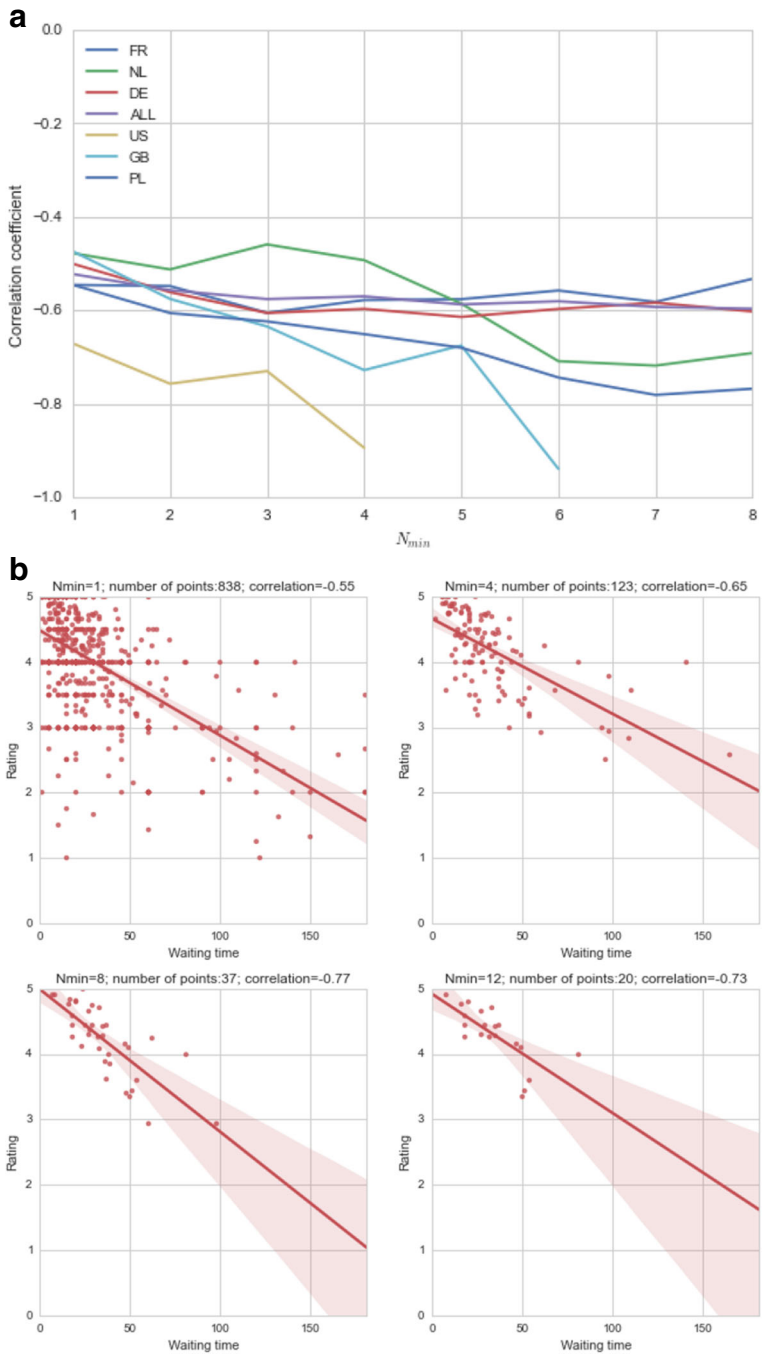

Fig. 1 a Correlation coefficient between waiting time and rating for all countries. **b** Correlation between waiting time and rating for different N_{min} for Poland

Road type analysis

Coordinates of a hitchhiking location, do not reveal anything specific about its relative emplacement. Obviously, hitchhiking points are located next to roads, but it is unknown where exactly: either on a motorway or a small village road. Therefore, a question of assigning the point to a road becomes crucial. A natural criteria is to classify roads by their type using road hierarchy [25]. Therefore the road type of a point is determined by the type of its closest road.

To implement our experiments, we use *OpenStreetMap* and its road hierarchy [26]. For example, the category

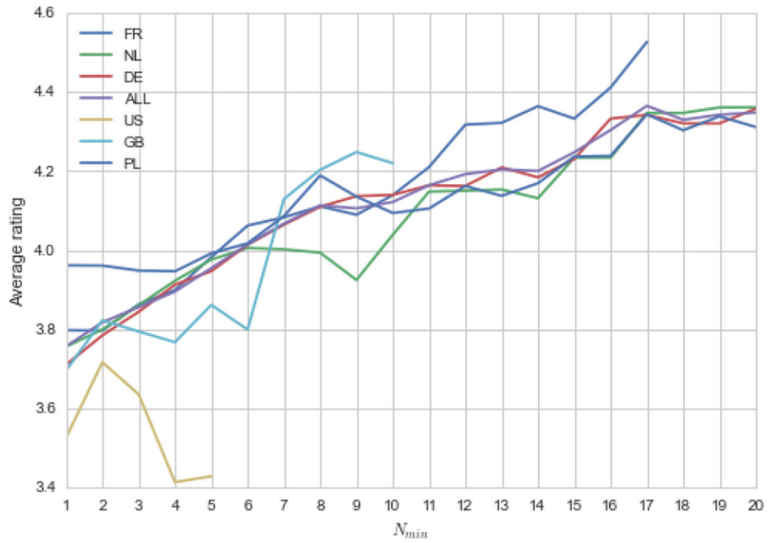

Fig. 2 Average rating and N_{min}

motorway is the highest among 6 main categories and *unclassified* is the lowest, while *trunk* is related to link between roads of different categories, i.e. ramps. The same classification hierarchy of roads may vary between countries. Thus, some different categories of roads may be represented in a different way, so to make our analysis feasible, we need to research roads from one country. In the Table 3, we illustrate the ratio of how many points of each country dataset were assigned to each type of road, how many roads there are in *OpenStreetMap* in that country. In this case, if a proportion is close to zero, than roads of this type are not popular for hitchhikers. The greater the proportion is, the more popular the road type is.

Link roads are especially popular for hitchhikers, followed by motorway and trunk roads. In addition, tertiary and unclassified roads are not popular, even though they are the most common types of roads in all countries. Note that this data is very specific to countries: for example, Polish hitchhikers tend to use primary roads instead of motorways, while trunk roads are especially popular in Great Britain. Therefore, recommendations for the desired application should consider individual features of each country.

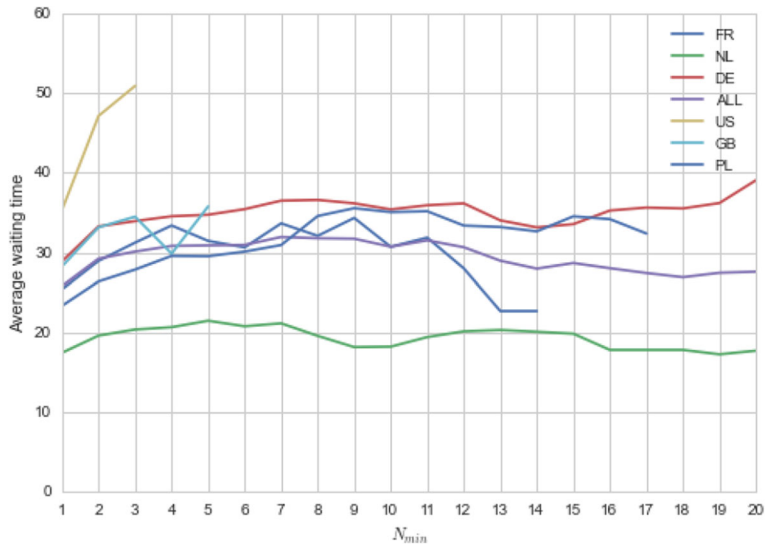

Fig. 3 N_{min} and waiting time and rating

Table 3 Road types distribution by country

Road type	France	Netherlands	Germany	USA	Great Britain	Poland
Motorway	3.42	2.30	1.93	0.49	0.92	0.97
Trunk	6.68	3.67	1.15	1.70	5.31	0.70
Primary	6.23	9.60	2.53	1.40	2.58	6.77
Secondary	1.12	4.91	0.62	0.51	0.51	1.30
Tertiary	0.15	0.28	0.20	0.19	0.12	0.20
Unclassified	0.06	0.10	0.21	0.19	0.14	0.04
Motorway_link	49.77	17.64	12.57	11.55	10.96	15.05
Trunk_link	48.47	13.63	7.15	3.19	4.68	5.80
Primary_link	13.01	7.24	4.27	2.10	2.00	29.20
Secondary_link	17.86	5.61	3.58	2.15	0.00	9.72
Tertiary_link	6.94	0.00	3.64	0.00	5.15	2.77

After the popularity of road types, we investigate ratings and waiting times at different road types. In the Fig 4a and 4b average rating and waiting time are shown in respect to N_{min}. We did not include roads secondary links and tertiary links, which are closest to 2 and 0 points respectively. Waiting time is different at different road types. For example, on primary roads it is almost half than on a motorway. In addition, motorway link roads have almost 50% less waiting time than motorway itself, which is important for long-distance travellers: they

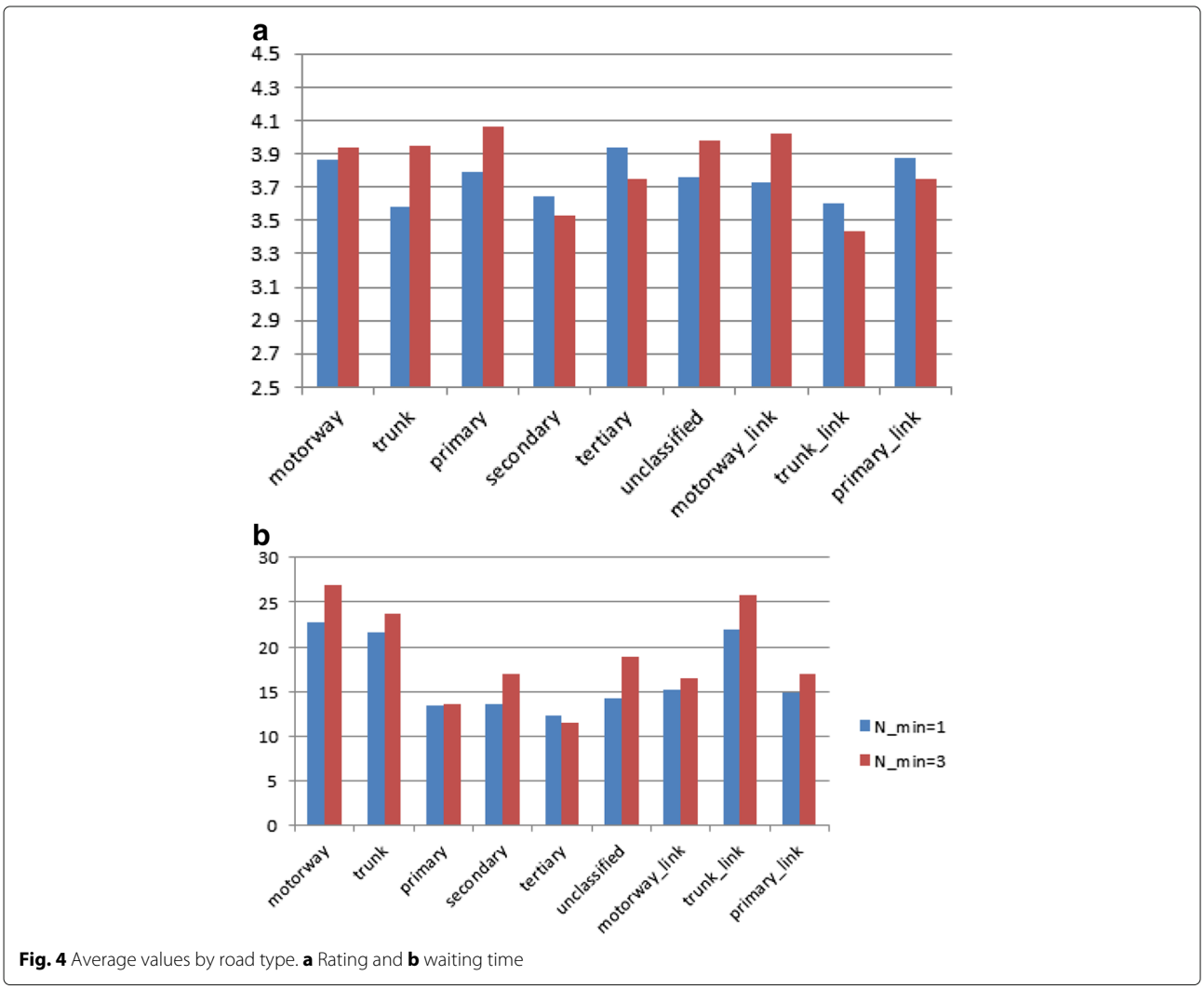

Fig. 4 Average values by road type. **a** Rating and **b** waiting time

should look for a proper position on a ramp instead of a motorway itself. Potentially, it gives us an opportunity to group points by road types: high-speed roads (motorway, trunk) or low-speed roads(others), while slip roads may constitute the third group. In terms of rating, trunk link roads and secondary roads in average get smaller ratings than other types of roads. For the trunk links, it may be related to higher waiting time, while secondary roads may seem inefficient due to the properties of traffic on them. For example, there could be many cars stopping, most of which do not go far.

Feature analysis

Following the verbal descriptions of good hitchhiking locations described in "Methodology" section, we crawled the features that are usually located next to roads: gas station, bus stop, traffic light, restaurant, parking. These features are extracted from *OpenStreetMap*.

To begin with, we may find what is the distance from *Hitchwiki* dataset points to these features. For each hitchhiking location, we assign distances to the each closest feature. The histograms of distributions are given in the Fig. 5.

Following, we investigate relationships between features, waiting times and ratings of locations in Fig. 6a. They depict the difference in waiting times/ratings between points that are located in features neighbourhoods to the points that are located far from them. For example, if the waiting time difference for bus stop feature at distance 0.02 km is -9 min, it means that average waiting time for points which have the closest bus stations less than 20 m away is 9 mins less than the rest of points.

Since we have many attributes derived from different sources, multicollinearity becomes an important question, i.e. when one or more attributes are highly correlated. For example, in many cases bus stops are located near traffic lights, so distances to these facilities may be somehow correlated. To measure it, we take a subset of points for $N_{min} = 3$ and find variance inflation factors (VIF) for the attributes for each country. VIF is a common

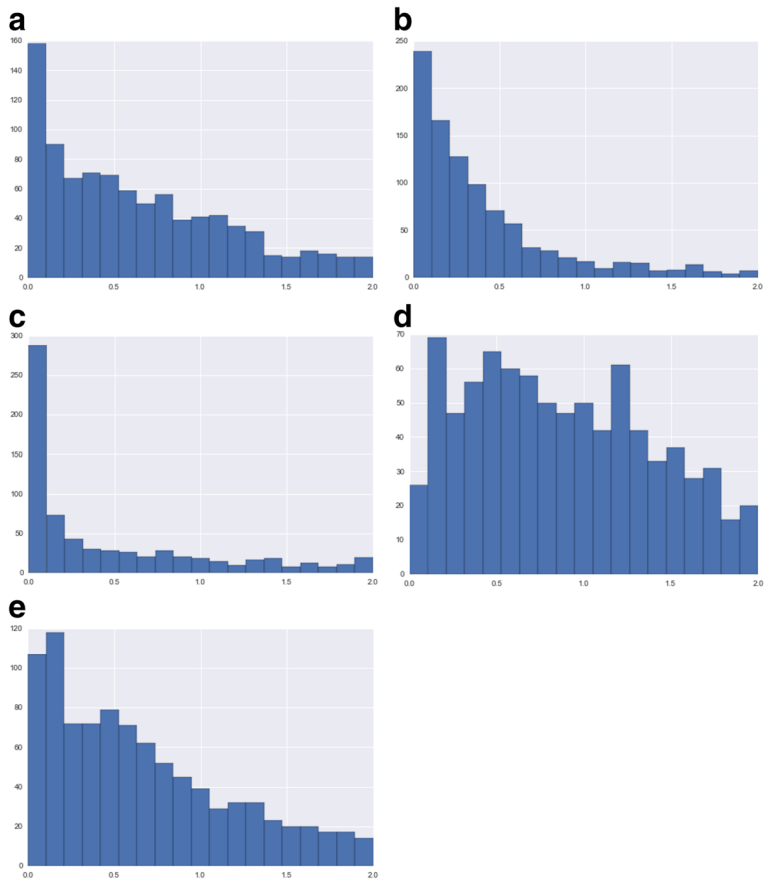

Fig. 5 Histogram of distance from dataset points to closest facilities. X-axis: distance to the closest facility in km, y-axis: number of hitchhiking locations with this distance. **a** gas station, **b** bus stop, **c** traffic light, **d** restaurant and **e** parking

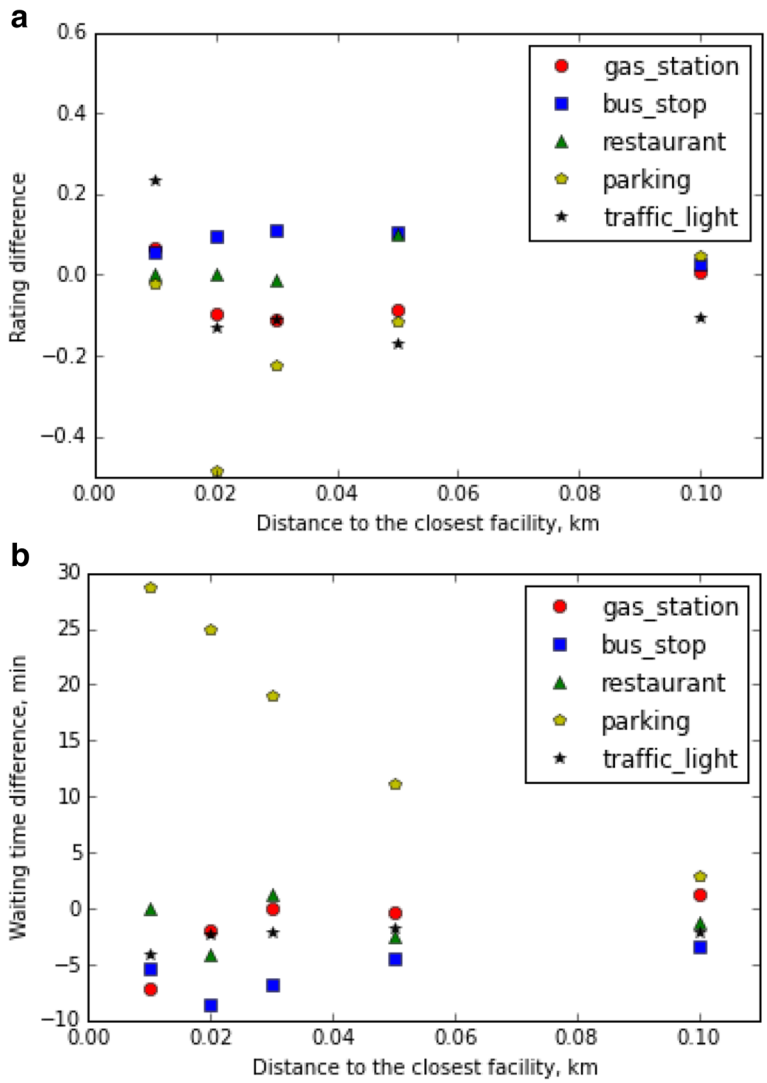

Fig. 6 Difference in waiting time/rating between points which are close to features neighbourhoods vs those which are not. **a** Rating difference and **b** waiting time difference

value which measures how much the variance of the estimated regression coefficients are inflated as compared to when the predictor variables are not linearly related, and it assumes that multicollinearity is high when VIF are larger than 5.

Variance inflation factors are given Table 4. We see that the VIF larger than 5 appears only for USA, the smallest dataset. In this case, the correlation comes from the fact that one of the road types appears only twice in the whole US dataset. After removing this column, VIF becomes 3.94.

However, our main goal is to have a system that predicts location rating, and in this case multicollinearity is not related to its performance, rather to coefficients and respective standard errors of attributes in a linear model. In addition, Random Forest can handle multicollinearity due to probabilistic sampling of a set of attributes. These results will be presented in the following section.

Another important statistics is p-value for each attribute, which tests the null hypothesis that the

Table 4 Variance inflation factors for different countries

Country	No. of points	Maximum VIF
Germany	799	1.806
France	513	2.194
Poland	332	3.089
Netherlands	233	3.678
Great Britain	156	2.539
USA	49	21.204

coefficient has no effect (equal to zero). Therefore, low p-values favor the alternative hypothesis. In our case, we measure p-values for the same data for the 3 biggest countries, and select attributes with p-values less than 0.05. The results are given in Table 5, and only_hw corresponds to a feature of point being near a highway without links nearby. For the other 3 countries, there is not enough data to have p-values small for the attributes.

To sum up, the most important features are gas stations, bus stops and traffic lights, and high proportion of points are located next to them. All of those generally improve waiting times and ratings. The more the distant is, the more random the data becomes, therefore all time and rating differences converge to zero. Even though some hitchhiking locations are located next to parkings, usually those points may have higher waiting times and lower rantings. Therefore, this facility is not recommended to hitchhike at, probably due to a low number of cars passing by parkings. However, the usage of these features in hitchhiking recommender system should be adjusted to specific countries.

Classification and regression

Since the main interest lies in distinguishing efficient and non-efficient points for a hitchhiker recommender system, the natural idea is to use classification of good and bad hitchhiking points for this purpose. For the following experiments, data from all countries will be used according to different values of N_{min}. The idea is to train the classification algorithm to distinguish points with high and low ratings based on the given list of attributes: type of road, distances to the closest gas station, bus stop, traffic light, and also if the point is around a highway link, around a highway without link, is it isolated from all facilities. First part of them has rating more than 4, and the rest have rating less than 3, and they correspond to 2 classes: "good" and "bad" points.

Accuracy of classification algorithms based on the given list of features for the classes of points with low and high ratings are given in the Fig. 7, in a setting when there are 66% random points selected for training and the rest 34% for testing. As it was mentioned above, as N_{min} is increasing, the average rating is increasing, so there are less points with low ratings, and the splitting into training and testing data is done after filtering. Therefore, the low rating class is oversampled in the experiments, and therefore we calculate the error bars for each of the classifier based on a sample of 1000 experiments.

In addition, we present the results of regression model to estimate location rating without division into two classes. In this case, there is no problem of imbalanced classes. The results of regression models are presented in Fig. 8.

To conclude, the results show high accuracy of given set of attributes for classification of efficient and not-efficient points for hitchhiking, and the accuracy increases when points have been ranked more than once. Even though the classes are skewed due to the reasons mentioned above, average accuracy of KNN and Random forest classifiers achieves 75% for special countries. Classification graphs for countries have similar structure, but results may vary more due to the less points in dataset, and since more uncertainty. The same feature set is also feasible for the regression problem, and the linear regression model achieves the best performance on median absolute error of 0.3 rating points.

Table 5 Attributes with low p-values

Country	No. of points	Attributes with p-values <0.05
Germany	799	secondary, unclassified, trunk_link
France	513	primary_link, motorway_link, residential, secondary, primary.
Poland	332	motorway, trunk, only_hw

Text analysis

Subsequent analysis is related to descriptions and comments of locations in the dataset. Here we try to make the first stage of text analysis, targeting to identify the relationships between keywords and attributes they represent. For this analysis, we also need to use subset from 1 country, because verbal descriptions of roads may be different for different countries, so we discuss the results for the Netherlands. For example, in this country roads starting with A (like A5, A19 etc.) correspond to largest motorways, while N-roads are used for general roads connecting towns. For each road type, we calculate how many points corresponding to this road type (e.g. their closest road is a road of specific road type from hierarchy) include a particular keyword. After that, for each keyword we have feature vector of frequency of its usage in each of these road types. For example, if a keyword "Gas station" is used in 50% of motorways and 5% of slip roads, its feature vector will be [0.5, 0.05] assuming there are only 2 road types. After that, the correlation matrix is computed to estimate the similarity between each pair of keywords, which finds the relationships between different keywords. The more the correlation coefficient is, the more similar these terms are.

Pearson correlation between keywords is given in Table 6.

This analysis proves the fact that that synonyms (as gas/petrol station, ramp/slip) have very high similarity, and also provides some valuable insights. For example,

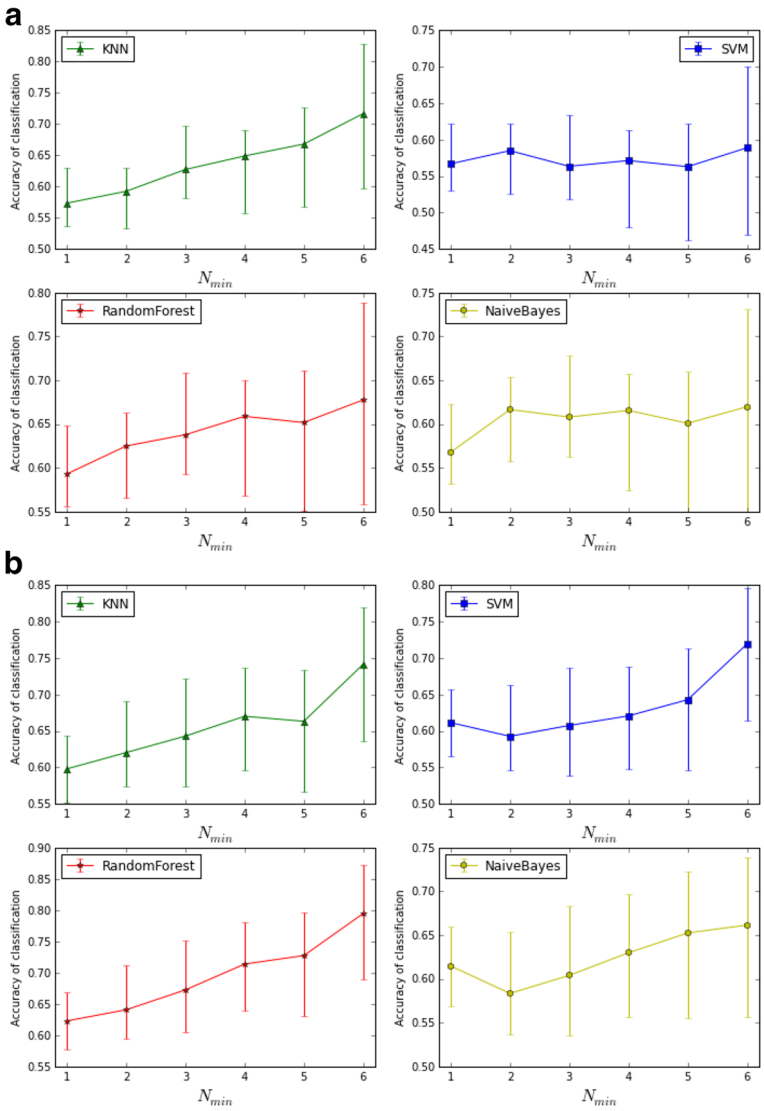

Fig. 7 Accuracy of classifiers. **a** Points in all countries and **b** points in Germany

bus stops are highly associated with N-type roads, while gas stations are more popular on motorways. However, as it was mentioned before, usage of keywords is mostly limited to each country, with corresponding keywords (names of roads etc.). Another complication is arising because in some countries comments are given not in English. For the future system, the information extracted from comments may be included into the recommender system.

Conclusions

In this work, we addressed a problem of finding important factors for hitchhiking locations, done a comprehensive analysis of *Hitchwiki* dataset of over 15,000 hitchhiking experiences. We found out correlation between rating and waiting time, but only if locations were assessed a few times. Moreover, the number of assessment provides more credibility to the assessment but also makes analysis harder due to the sparseness of the data. Properties of different road types were examined, and regional roads linking towns in average have two times less waiting times than motorways as well as motorway ramps. Generally, link roads and motorways are the most popular comparing to their relative size in countries' road network, while urban roads and small roads, which make the most of road network, attract much less hitchhikers. However, these values vary from country to country due to local features.

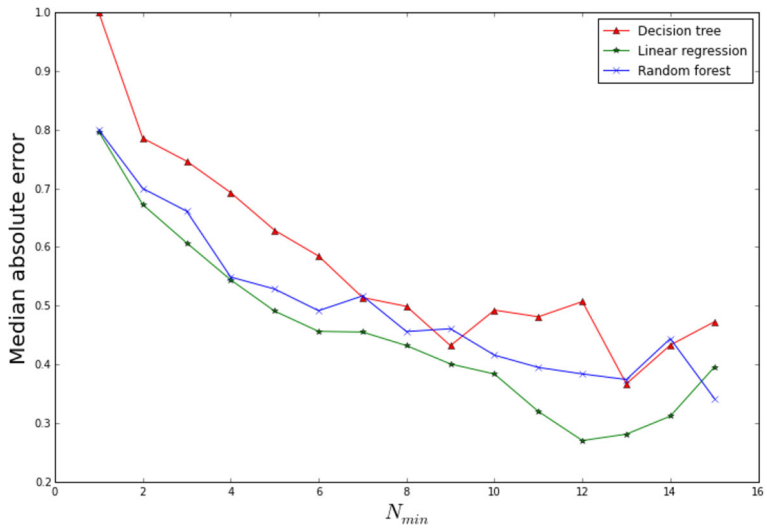

Fig. 8 Performance of regression models

Most important road facilities for hitchhikers include gas stations, bus stops and traffic lights, so the majority of hitchhiking locations are located next to them. Closeness to them generally improves waiting times and ratings. Parking spots indeed decrease hitchhiking performance, and thus need to be avoided. We also analysed comments and descriptions to the locations, and proved the applicability of our method due to high similarity values between synonyms. Some valuable insights include popularity of gas stations on motorways and bus stops at special types of regional roads. Finally, our feature set has reached 76% accuracy in classification tests and 0.3 median absolute error for the regression model, which is the first result made for analysis of hitchhiking datasets.

The *Hitchwiki* dataset itself lacks information about some details of hitchhiking travel experience, most importantly the time and destination of trips. Hitchhiking at daytime and night may be completely different, as well as asking a short ride may always imply less waiting time than finding a driver for a long-distance trip. However, even without these parameters, we answered the stated questions and identified the best locations for hitchhiking.

Considering future works, this research benefits to developing an app to connect hitchhikers and drivers as well as assist hitchhikers at all stages of trips, including route planning and location selection. For example, having an initial destination and source as input, the app should select the optimal route and help to find proper locations on the way, which has been done by implementing a classification model. Depending on the system design, regression models may be used instead of classification. In addition, provided feedback (trip description, waiting and ride time) in the app will improve security and accuracy of predictions.

Table 6 Correlation between different words according to types of roads they describe

	Petrol station	A[1, 2...]	Ramp	Motorway	N[1, 2...]	Bus stop	Slip	Gas station	Highway
Petrol station	1	0.82	0.7	0.98	0.3	0.42	0.72	0.95	0.8
A[1, 2...]	0.82	1	0.97	0.83	0.58	0.66	0.98	0.83	0.97
Ramp	0.7	0.97	1	0.72	0.63	0.67	0.98	0.72	0.94
Motorway	0.98	0.83	0.72	1	0.37	0.47	0.74	0.96	0.82
N[1, 2...]	0.3	0.58	0.63	0.37	1	0.97	0.53	0.51	0.68
Bus stop	0.42	0.66	0.67	0.47	0.97	1	0.59	0.62	0.77
Slip	0.72	0.98	0.98	0.74	0.53	0.59	1	0.72	0.94
Gas station	0.95	0.83	0.72	0.96	0.51	0.62	0.72	1	0.87
Highway	0.8	0.97	0.94	0.82	0.68	0.77	0.94	0.87	1

Additional file

Additional file 1: Additional graphs. (PDF 1540 kb)

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
OV carried out writing and done experiments. LK and KR contributed in writing, analysis, and presentation of results. All authors read and approved the final manuscript.

Received: 22 August 2016 Accepted: 20 September 2016
Published online: 19 December 2016

References
1. Wikipedia: Hitchhiking. https://en.wikipedia.org/wiki/Hitchhiking. Accessed May 2016.
2. Nederlandlift: Survey About Hitchhiking in the Netherlands. http://www.nederlandlift.nl/nieuws/results-survey-hitchhiking-netherlands/ Accessed May 2016.
3. Wikitravel: Tips for Hitchhiking. http://wikitravel.org/en/Tips_for_hitchhiking. Accessed May 2016.
4. Chesters G, Smith D. The Neglected Art of Hitch-hiking: Risk, Trust and Sustainability. Bull Psychon Soc. 1975;5(6):459–461. Sociological Research Online. (2001). 00014.
5. Morgan CJ, Lockard JS, Fahrenbruch CE, Smith JL. Hitchhiking: Social signals at a distance. Bull Psychon Soc. 2013;5(6):459–61. doi:10.3758/BF03333299.00018. Accessed 31 Aug 2015.
6. Guéguen N, Lamy L. Hitchhiking women's hair color. Percept Mot Skills. 2009;109(3):941–8. doi:10.2466/pms.109.3.941-948.00001.
7. Guéguen N. Bust size and hitchhiking: a field study. Percept Mot Skills. 2007;105(3f):1294–8. doi:10.2466/pms.105.4.1294-1298.00000. Accessed 14 Dec 2015.
8. Guéguen N, Lamy L. The effect of facial makeup on the frequency of drivers stopping for hitchhikers. Psychol Rep. 2013;113(1):1109–13. 00001.
9. Guéguen N. Effect of Humor on Hitchhiking: A Field Experiment. North Am J Psychol. 2001;3:00008.
10. Guéguen N, Stefan J. Hitchhiking and the 'sunshine driver': further effects of weather conditions on helping behavior'. Psychol Rep. 2013;113(3):994–1000. doi:10.2466/17.07.PR0.113x30z8.00001.
11. Pazhoohi F, Burriss RP. Hijab and "Hitchhiking": A Field Study. Evol Psychol Sci. 2015: 1–6. doi:10.1007/s40806-015-0033-5.00001. Accessed 14 Dec 2015.
12. Verhart F. Hitchhiking Stats. http://hitchwiki.org/en/User:Fverhart. Accessed May 2016.
13. Ausfahrtlinks: Autostoppen in Österreich. https://ausfahrtlinks.wordpress.com/. Accessed May 2016.
14. Furuhata M, Dessouky M, Ordóñez F, Brunet ME, Wang X, Koenig S. Ridesharing: The state-of-the-art and future directions. Transportation Res Part B: Methodol. 2013;57:28–46. doi:10.1016/j.trb.2013.08.012.00068. Accessed 24 Aug 2015.
15. Chan N, Shaheen S. Ridesharing in North America: Past, Present, and Future. Trans Rev. 2012;32(1):93–112. 00098. http://www.tandfonline.com/doi/abs/10.1080/01441647.2011.621557.
16. Ma S, Wolfson O. Analysis and Evaluation of the Slugging Form of Ridesharing. In: Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems SIGSPATIAL'13. New York: ACM; 2013. p. 64–73. doi:10.1145/2525314.2525365.00012 . http://doi.acm.org/10.1145/2525314.2525365. Accessed 17 Aug 2015.
17. Yao X. Where are public transit needed – Examining potential demand for public transit for commuting trips. Comput Environ Urban Syst. 2007;31(5):535–50. doi:10.1016/j.compenvurbsys.2007.08.005.00031. Accessed 09 Mar 2016.
18. Hitchlog: Hitchhiking Logs. http://www.hitchlog.com/. Accessed May 2016.
19. HitchWiki: Hitchhiking Maps and Wiki. http://hitchwiki.org/. Accessed May 2016.
20. Hopista: Findings About Hitchhiking. http://www.hopista.org/findings/. Accessed May 2016.
21. Bouchard AM. Hitchhiking. http://hitchwiki.org/en/images/en/c/cb/Hitchhiking_-_Neo-nomad.pdf. Accessed May 2016.
22. Senaratne H, Mobasheri A, Ali AL, Capineri C, Haklay MM. A review of volunteered geographic information quality assessment methods. Int J Geograph Inform Sci. 20161–29. doi:10.1080/13658816.2016.1189556.00004 . Accessed 20 Oct 2016.
23. Jiang S, Alves A, Rodrigues F, Ferreira J, Pereira FC. Mining point-of-interest data from social networks for urban land use classification and disaggregation. Comput Environ Urban Syst. 2015;53:36–46. doi:10.1016/j.compenvurbsys.2014.12.001.00003 . Accessed 09 Mar 2016.
24. Sun Y, Fan H, Bakillah M, Zipf A. Road-based travel recommendation using geo-tagged images. Comput Environ Urban Syst. 2015;53:110–22. doi:10.1016/j.compenvurbsys.2013.07.006.00022 . Accessed 09 Mar 2016.
25. Brindle RE. Road Hierarchy and Functional Classification (1989). 1996, p. 00017. http://trid.trb.org/view.aspx?id=1205239. Accessed 15 Dec 2015.
26. OpenStreetMap: Road Hierarchy. 2016. http://wiki.openstreetmap.org/wiki/Key:highway. Accessed May 2016.