Data in brief 25 (2019) 104039

Contents lists available at ScienceDirect

Data in brief

journal homepage: www.elsevier.com/locate/dib

Data Article

On-street parking availaibilty data in San Francisco, from stationary sensors and high-mileage probe vehicles

Fabian Bock [a], Sergio Di Martino [b, *]

[a] Institute of Cartography and Geoinformatics, Leibniz University, Hannover, Germany
[b] Department of Electrical Engineering and Information Technologies, University of Naples “Federico II”, Naples, Italy

ARTICLE INFO

Article history:
Received 29 November 2018
Received in revised form 9 May 2019
Accepted 15 May 2019
Available online 4 June 2019

ABSTRACT

This dataset contains records of the measured on-street parking availability in San Francisco, obtained from the public API of the *SFpark* project.[1]

In 2011, the San Francisco Municipal Transportation Agency (SFMTA) started a project on smart parking, called *SFpark*, whose goal was the improvement of on-street parking management in San Francisco, mostly by means of demand-responsive price adjustments [1]. One of the key points of the project was the collection of information about on-street parking availability. To this aim, about 8,000 parking spaces were equipped with specific sensors in the asphalt, periodically broadcasting availability information. The SFpark project made available a public REST API, returning the number of free parking spaces and total number of provided parking spaces per road segment, for 5,314 parking spaces on 579 road segments in the pilot area. We collected parking availability data from 2013/06/13 until 2013/07/24, by querying this API at approximately 5-min intervals. As a result, we obtained in total about 7 million observations of parking availability on the road segments. These observations represent the first dataset we are providing.

In addition, we simulated the achievable sensing coverage of on-street parking availability that could be achieved by a fleet of taxis, if they were equipped with sensors able to detect free parking spaces, like side-scanning ultrasonic sensors [3], or

* Corresponding author.
E-mail address: sergio.dimartino@unina.it (S. Di Martino).

[1] http://sfpark.org/.

https://doi.org/10.1016/j.dib.2019.104039

windshield-mounted cameras [4]. In particular, by exploiting real taxi trajectories in San Francisco from the Cabspotting project [5], we first computed the frequencies of taxi visits for each road segment covered by the SFpark sensors. Then, we downsampled the first dataset, in order to have a parking availability information for a road segment at a given time only in presence of a transit of a taxi on that segment at that time. This step was replicated for 5 different sizes of taxi fleets, namely 100, 200, 300, 400, and 486. Consequently, in total six datasets are available for further research in the field of on-street parking dynamics.

All these datasets can be downloaded at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YLWCSU.

Specifications Table

Subject area	*Social Sciences: Transportation*
More specific subject area	*Smart parking, On-street parking, Crowd sensing, Probe vehicles.*
Type of data	*Text Files, in Comma Separated Value format, containing Parking Availability Data.*
How data was acquired	*A first dataset has been acquired by means of the SFpark API, reporting parking availability data collected from stationary sensors in the asphalt. The other datasets are a subset of the first one, obtained by simulating the crowd-sensing capabilities of a fleet of 100, 200, 300, 400 and 486 taxis, used as probe vehicles for parking availability sensing.*
Data format	*Raw + Downsampling of raw, obtained by simulation.*
Experimental factors	*GPS data from the taxis were map-matched on the road network provided by Open Street Map. Implausible GPS points were discarded.*
Experimental features	*For 420 road segments in San Francisco, there is a record every 5 minutes containing information on the total and free number of parking stalls.*
Data source location	*San Francisco, USA.*
Data accessibility	*Data is available in the Harvard DataVerse Repository. URL: https://dataverse.harvard.edu /dataset.xhtml?persistentId=doi:10.7910/DVN/YLWCSU*
Related research article	*Smart Parking: Using a Crowd of Taxis to Sense On-Street Parking Space Availability.* *In IEEE Transactions on Intelligent Transportation Systems (2019)* *DOI: 10.1109/TITS.2019.2899149* [1].

Value of the data

- The first open-access dataset reporting real-world on-street parking availability data on a wide urban area.
- Our dataset provides real on-street parking availability data collected from stationary sensors in the asphalt, for about 500 road segments in San Francisco downtown, plus simulated availability data that would have been collected by a fleet of taxis, acting also as probe vehicles.
- Researched from many fields (ICT, Transportation, Social Sciences, etc.) can exploit this dataset to investigate, for the first time in a data-driven fashion, the dynamics underlying on-street parking.
- Our dataset has a fine-grained temporal resolution, of 5 minutes, allowing for detailed investigations.
- This is the first and only dataset to investigate the potentialities of probe vehicles to sense on-street parking data. The information about probe vehicle sensing are built on real trajectories from 100, 200, 300, 400 and 486 taxis.

1. Data

The data files we provide in this article are in CSV format. Data have been collected according to the pipeline described in Fig. 1, for five different sizes of the fleets of taxi. The size of the fleet is the suffix at the end of the file name (e.g. XXX_100taxis.tab is the set of files obtained by a simulated fleet of 100 vehicles). The dataset contains a set of six files, as described in the following table:

Filename	Description
sfpark_filtered_segments.csv	Geometrical information on the road segments under investigation
sfpark_filtered_XXXtaxis.csv	Parking availability information for the investigated road segments, as may have been sensed by a fleet of XXX taxis (where XXX can be 100, 200, 300, 400, or 486).

1.1. *Format of the sfpark_filtered_segments.csv*

Each line of the file corresponds to a road segment with on-street parking stalls, covered in the SFpark pilot area (also called block face by the SFpark project). The parking segments of the SFpark pilot area were filtered for data plausibility as described in Refs. [1,7]. The columns correspond to the following content:

- *segmentid:* ID of the parking segment (also called *block face* by SFpark) as defined by the SFpark project.
- *streetname*: Name of road with house number range.
- *startx, starty, endx, endy*: WGS84 coordinates of start and end point of parking segment.

1.2. *Format of the sfpark_filtered_XXXtaxis.csv*

Each line of the files corresponds to the parking situation of a parking segment described in the previous file, at a specific timestamp. The parking segments of the SFpark pilot area were filtered for data plausibility, as describe in Ref. [1]. The crowd-sensing observations were computed as described in Ref. [7].

The columns correspond to the following content:

- *timestamp*: Timestamp reporting when the SFPark API was polled, rounded to the closest minute.
- *segmentid*: ID of the parking segment as defined by SFpark, whose geometry is described in the file *sfpark_filtered_segments.csv.*
- *capacity*: current total number of parking spaces in the parking segment (please note that the capacity may vary over time, due to parking restrictions)
- *occupied*: current number of occupied parking spaces in the parking segment, reported by the SFpark sensors.

Then there is the block of simulated availability information, as they might have been sensed by the probe vehicles for parking crowd-sensing. Since we downsampled the original dataset of trajectories (please refer to section 2 for further details), we report data about 10 repetitions, to reduce biases due to the random downsampling, as follows:

- *observedK*: has value 1 if, in the time frame indicated by the timestamp, at least one taxi passed over this road segment, 0 otherwise. *K* corresponds to the K^{th} repetition of the random sampling of taxis and of their lane choice. For any further use of this dataset, whether *observedK* has value 1, this means that a probe vehicle would have reported the same value of *occupied* as by the sensors of SFpark.
- *diffK*: difference between the last occupancy measurement of the taxis and the current number of occupied parking spaces; K like above.

2. Experimental design, materials and methods

The pipeline to collect the experimental datasets is described in Fig. 1.

2.1. Data from stationary sensors

The real-world on-street parking availability data comes from the *SFpark* project [2]. This project made available a public REST API, returning the number of free parking spaces and total number of provided parking spaces per road segment, for 5,314 parking spaces on 579 road segments in the pilot area. We collected parking availability data from 2013/06/13 until 2013/07/24, by querying this API at approximately 5-min intervals. As a result, we obtained in total about 7 million observations of parking availability on the road segments. Let us observe that, due to design of the *SFpark* API, it was not possible to collect availability information at a stall granularity, but rather at road segment level.

To cleanse the dataset from implausible sensor values, we excluded a road segment from our investigation if, for more than three days, data was missing or the number of occupied parking spaces on it was constant. We removed road segments never showing an occupancy rate higher than 85%, too, based on the assumption that either some sensors failed, or parking was not competitive there and that road segment was thus less relevant to monitor. Finally, we excluded tow-away periods, such as street cleaning or peak hour drive way periods, using the *SFpark* API, which reported, during those time frames, both zero parking capacity and zero occupied spaces for the involved road segments. As a result, of the original 579 road segments, we kept in our dataset 420 of them, for a total of more than 5 million observations. To have a common geo-spatial layer, we used the OpenStreetMap road network. All these steps are represented in block (1) of Fig. 1.

2.2. Simulated crowd-sensed data

The simulation of the obtainable crowd-sensed parking availability data is based on the hypothesis that all the considered probe vehicles are equipped with sensors able to detect empty parking spots while passing by a road segment, like described in [3] or [4], and sending this information to a back-end infrastructure, where it is aggregated to obtain a dynamic parking availability map.

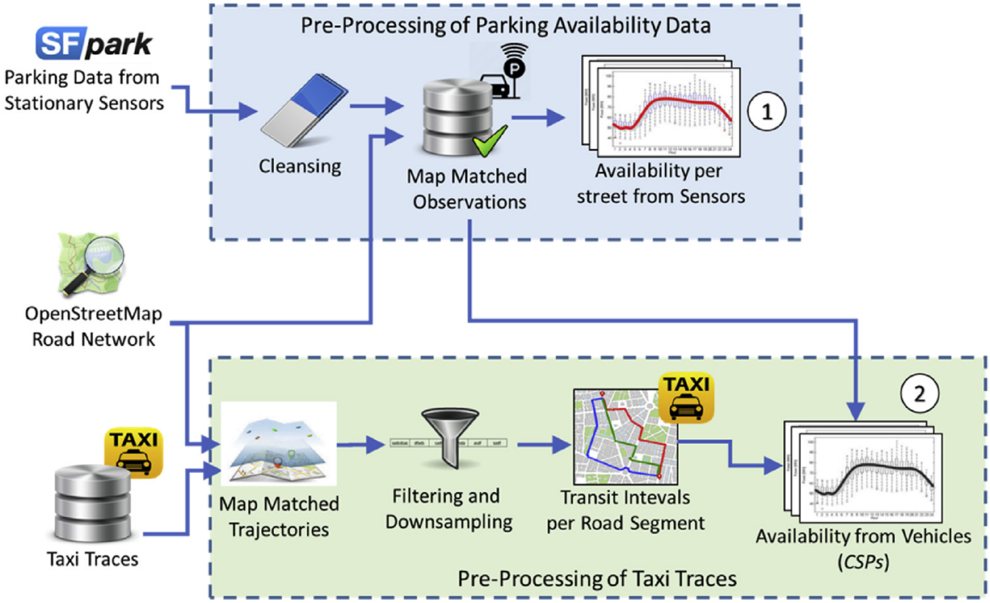
SFpark
Parking Data from Stationary Sensors
Pre-Processing of Parking Availability Data
Cleansing
Map Matched Observations
Availability per street from Sensors
1
OpenStreetMap Road Network
TAXI
Taxi Traces
Map Matched Trajectories
Filtering and Downsampling
Transit Intevals per Road Segment
2
Availability from Vehicles (CSPs)
Pre-Processing of Taxi Traces

Fig. 1. Schematic view of the pipeline used to build the datasets.

To simulate the achievable spatio-temporal coverage of crowd-sensed parking data, we used a public repository of taxi trajectories in San Francisco, to measure how often a road segment might be visited by a taxi, sensing its parking availability.

To this aim, we started by exploiting a repository of taxi traces was collected within the *Cabspotting* project [5]. Each taxi periodically sent to a central server a Floating Car Data (FCD) record, containing information on its latitude, longitude, timestamp, and passenger occupancy. The resulting dataset contains 11,219,955 observations, collected from 536 vehicles of the Yellow Cab company, over 25 days in the San Francisco Bay Area.

We first map-matched the sequences of GPS points with the OpenStreetMap road network, using the approach described in Ref. [6]. Then, we split each taxi's FCD flow into a set of independent trajectories. The trajectories corresponding to a total of 8,839,942 GPS points (78.8% of the original dataset) were retained in the considered dataset. About one third of the missing points were discarded due to too few GPS points per trip. The remaining two thirds were excluded because at least one segment of the trajectory required an implausible speed, with respect to the underlying road network, mainly due to sporadic GPS sensing errors. As next step, a spatial filtering of the trajectories was performed, by removing all the trajectories not overlapping road segments covered by the *SFpark* project.

In order to simulate different sizes of the taxi fleet, from the full dataset of 486 taxis, a downsampling for a lower number N of taxis (100, 200, 300, 400) was computed, by randomly sampling vehicles. To reduce selection biases, we generated ten independent subsamples for each considered fleet size. The data we are providing is computed on the average over these ten subsets.

Since only probe vehicles moving on a lane adjacent to a parking lane can monitor availability, the number of lanes and the driving direction must be considered. As the lane choice of the taxi drivers is unknown, we considered a uniform distribution of the taxis over the lanes, and therefore, only 1/#(lanes) of the visits (randomly chosen for each subsample) in the corresponding driving direction were used for the subsequent calculations.

To compute the spatio-temporal sensing coverage of the taxi fleets, it is necessary to compute the temporal resolution of the taxi visits for each road segment, as described in Ref. [7]. The average time among two subsequent taxi visits for each road segment were computed also for smaller fleets of 50, 100, 200, 300 and 400 vehicles.

Finally, to generate the dataset of crowd-sensed parking availability for a fleet of size *s* (*CSPs*), for each road segment and hour of each day of the week, we used the taxi transits of this hour to mask observations from the original *SFpark* dataset. More in detail, each *CSPs* dataset, represented in block (2) of Fig. 1, has the same number of road segments and the same temporal resolution of the *SFpark* one, namely 420 segments with a record every 5-min, and is generated as follows: for each road segment r and each 5 minutes time frame, centered on an observation in the *SFpark* dataset (−2.5min and +2.5min), we checked the visit of at least one probe vehicle from the fleet of size s. If this is the case, a taxi visit function $V_{taxi}(s, r, t)$ yields 1, else it yields 0. If $V_{taxi}(s, r, t)=1$, we simulated the crowd-sensing of parking availability by copying the corresponding information from the *SFpark* dataset into *CSPs*. On the contrary, if no vehicle visited the road segment r in the time frame t, no parking information could have been crowd-sensed, and the corresponding entry in the *CSPs* dataset is thus a missing observation.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Smart Parking: Using a crowd of taxis to sense on-street parking space availability, in: IEEE Transactions on Intelligent Transportation Systems, 2019, https://doi.org/10.1109/TITS.2019.2899149.

[2] SFMTA, SFpark: Putting Theory into Practice. Pilot Project Summary and Lessons Learned, 2014. http://sfpark.org/resources/docspilotsummary/.

[3] S. Mathur, T. Jin, N. Kasturirangan, J. Chandrasekaran, W. Xue, M. Gruteser, W. Trappe, Parknet: drive-by sensing of road-side parking statistics, in: Procs of the 8th International Conference on Mobile Systems, Applications, and Services. ACM, 2010, pp. 123–136.

[4] G. Grassi, P. Bahl, K. Jamieson, G. Pau, Parkmaster: an in–vehicle, edge–based video analytics service for detecting open parking spaces in urban environments, in: The Second ACM/IEEE Symposium on Edge Computing (SEC 2017), 2017.

[5] M. Piorkowski, N. Sarafijanovic-Djukic, M. Grossglauser, CRAWDAD Dataset Epfl/mobility (V. 2009-02-24), Feb. 2009. Downloaded from, http://crawdad.org/epfl/mobility/20090224.

[6] S. Axer, F. Pascucci, B. Friedrich, Estimation of traffic signal timing data and total delay for urban intersections based on lowfrequency floating car data, in: Procs. Of the 6th Mobil.TUM 2015, 2015.

[7] F. Bock, Y. Attanasio, S. Di Martino, Spatio-temporal road coverage of probe vehicles: a case study on crowd-sensing of parking availability with taxis, in: Societal Geo-Innovation, Ser. Lecture Notes in Geoinformation and Cartography, Springer International Publishing, 2017, pp. 165–184.