# DATA7: A Dataset for Assessing Resource and Application Management Solutions at the Edge

Emanuele Carlini
National Research Council
Pisa, Italy
emanuele.carlini@isti.cnr.it

Massimo Coppola
National Research Council
Pisa, Italy
massimo.coppola@isti.cnr.it

Patrizio Dazzi
University of Pisa
Pisa, Italy
patrizio.dazzi@unipi.it

Luca Ferrucci
University of Pisa
Pisa, Italy
luca.ferrucci@unipi.it

Hanna Kavalionak
National Research Council
Pisa, Italy
hanna.kavalionak@isti.cnr.it

Matteo Mordacchini
National Research Council
Pisa, Italy
matteo.mordacchini@iit.cnr.it

## ABSTRACT

This paper presents a dataset on edge devices and mobility patterns to comprehensively understand user behaviour and devices workload in Edge computing environments. The dataset is built on top of a publicly available dataset of cellular tower locations to simulate Edge devices, and on user mobility trajectories generated by a state-of-the-art simulator based on real location maps in the area of the city of Pisa, Italy. The resulting dataset reports the amount of vehicles in the range of about 200 Edge devices for each step of the simulation. The dataset can be used for various applications in edge computing and mobility, most notably for assessing results on resource and application management solutions at the edge in a realistic environment.

## CCS CONCEPTS

• **Networks** → *Network resources allocation*; • **Human-centered computing** → **Ubiquitous and mobile computing design and evaluation methods**; • **Information systems** → *Computing platforms.*

## KEYWORDS

user mobility, resource management, dataset, edge computing

**ACM Reference Format:**
Emanuele Carlini, Massimo Coppola, Patrizio Dazzi, Luca Ferrucci, Hanna Kavalionak, and Matteo Mordacchini. 2023. DATA7: A Dataset for Assessing Resource and Application Management Solutions at the Edge. In *Proceedings of the 3rd Workshop on Flexible Resource and Application Management on the Edge (FRAME '23), June 20, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589010.3595652

## 1 INTRODUCTION

In today's world, where mobile devices are prevalent, smartphones are gateways to a vast collection of data and applications. This significantly impacts the computing infrastructure that handles user requests passing through these gateways. Whilst the requests generated by the current devices already challenge the existing infrastructure, the envisioned future scenarios foresee even more impactful requests that are aimed at supporting highly-reactive multimodal human-computer interactions mediated through more advanced devices (e.g., next-generation HMDs).

Traditional cloud-based infrastructures do not represent a suitable solution for the next generation of mobile applications and devices due to the latency sensitiveness of applications and the data deluge generated by devices. To address such needs, a viable approach is to bring computing capacity near the end devices instead of transferring data and requests to remote computing infrastructures. Edge computing approaches the problem by exploiting edge data centres, which are a sort of "walking-sized" Clouds located pervasively in the environment (typically co-located with antennas for mobile communication [5, 21]), to perform computational tasks that would otherwise be processed remotely. Determining when and what edge data centres to exploit for a large set of users and applications while matching the requirements of the applications and the expectation of users is challenging. Even more, achieving it while ensuring optimal exploitation of resources is a hard task, as by the NP-completeness of the underlying problems. Many approaches to face the problem have been proposed so far [11, 14], however they usually assess their effectiveness only under specific conditions that are hardly repeatable or customisable (e.g., to assess the quality of the solution when conditions change).

To overcome these issues, this paper presents a comprehensive dataset on edge devices and mobility patterns that can be used for various applications in edge computing and mobility. Our approach uses the publicly available dataset of cellular tower locations [17] to use in combination with the user mobility patterns generated by the SUMO simulator [12] based on real location maps. The resulting artefact provides a more accurate and representative view of real-world scenarios, enabling researchers and practitioners to develop and evaluate edge computing and mobility solutions in a more realistic environment. Finally, to foster further research in the field, we made the dataset publicly available [4].

## 2 RELATED WORK

Using reference datasets to verify solutions' efficiency and effectiveness is common in computer science and engineering research.

*FRAME '23, June 20, 2023, Orlando, FL, USA*

ACM ISBN 979-8-4007-0164-1/23/06…$15.00
https://doi.org/10.1145/3589010.3595652

Specifically, this applies to validating resource and application management solutions in large computing infrastructures. There are notable examples of datasets that have been successful in this field. Some of such artefacts consist of publicly available real workload traces (such as Google cluster traces [8], Eucalyptus IaaS cloud Workload [6], and many others). The relevance of such data for the community is demonstrated by the large number of papers having these datasets as the main subject or at least as a primary entity for the evaluation activity [9, 10, 19, 20].

Other datasets provide collections of PoI relevant to edge computing research, such as antennas of mobile network carriers, properly organised for easing the investigation and comparison of different solutions. In this context, the dataset created within the BASMATI EU project [2] presents the Received Signal Strength Indicator (RSSI) of the smartphone WiFi transceivers of the attendants of a music festival on a specific day, as many WiFi access points recorded it. Among those, a prominent role for what concerns research on edge computing and user mobility is played by the artefact provided by Lai *et al.* [13]. Starting from such a dataset, many works have been developed[1].

A different but complementary approach has been followed by Xiang et al. [22]. In this work, the authors focus on networking aspects and release a dataset with data related to 3 randomly generated MEC topologies with increasing network size (from 25 to 100 nodes). Such topologies could be used to run extensive experiments and compare different solutions' performance concerning planning, scheduling, routing, etc. Alsaedi *et al.* [1] released a dataset called "TON_IoT Telemetry" includes Telemetry data of IoT-IIoT services, Operating Systems logs and Network traffic of IoT network collected from a realistic representation of a medium-scale network at the Cyber Range and IoT Labs at the UNSW Canberra (Australia). Starting from such a work, Zachos *et al.* [23] generate a set of IoT edge network-specific datasets based on the "ToN_IoT Telemetry" dataset.

All the datasets mentioned above aim to support research activities in edge computing or user mobility. However, no publicly available dataset supports validating and evaluating solutions considering user mobility and edge computing altogether.

## 3 THE DATASET

DATA7 is built by merging cellular towers' position with the movement of vehicles in the city of Pisa, Italy. The methodology we followed is depicted in Figure 1. We used two main sources of data: (i) synthetic vehicles trajectories made with SUMO to simulate the movement of active users, and (ii) the real position of cellular towers from OpenCelliD, which we consider a good approximation of the position of potential Edge computing nodes.

### 3.1 Sumo trajectories generation

Simulation of Urban MObility (SUMO) [3] is an open-source, microscopic, and continuous traffic simulator tool designed to handle a large set of traffic scenarios. SUMO networks consist of nodes and unidirectional edges representing street, waterways, tracks, etc. A single simulation scenario involves vehicles moving through a given road network. Each vehicle is modelled explicitly, has its own route, and moves individually through the network. SUMO networks include detailed information regarding possible movements at intersections and the corresponding right-of-way rules used to determine the dynamic simulation behavior. A useful tool for initial scenario preparation is the osmWebwizard [15] application, which has been used in this paper to generate the network of Pisa. It allows a simple selection of an area from a map display along with a set of parametrized traffic modes. The tool uses this information to download and import the network data from OpenStreetMap [7] with parameters corresponding to the selected traffic modes. Additionally, SUMO provides the SUMO-GUI [16] application (Figure 2), which allows observing the simulation in various aspects such as speeds, traffic densities, road elevation, or right-of-way rules. To evaluate a simulation scenario quantitatively, the simulation provides several output files, including: (i) Vehicle trajectories, (ii) traffic data, (iii) protocols of traffic light switching, (iv) traffic data aggregated for the whole simulation, and (v) Emissions and energy consumption.

To identify the network of routes in the Pisa city area and create a simulation scenario, we employed the SUMO OSMWebWizard tool, as illustrated in Figure 2. Our simulation scenario was generated with the "Import Public Transport" feature enabled, and we exported both busStops and trainStops. To generate car traffic, we configured the demand generation to include a *Through traffic control* of 4 and a *Count* of 8. We conducted a simulation comprising of 3600 steps, where each step in SUMO corresponds to one second by default. The Through Traffic Factor measures the likelihood of selecting an edge at the simulation area's boundary over an edge within the area. A higher value indicates more through traffic, with many vehicles departing and arriving at the boundary. The Count parameter defines how many vehicles are generated per hour and lane-kilometer [15]. In our simulation, we have not imposed any restrictions on the road selection.

### 3.2 OpenCelliD Preprocessing

To simulate the position of edge devices, we downloaded the location of cellular towers from the OpenCelliD website [17]. OpenCelliD collects information about cell towers in a community-driven fashion, i.e., data is primarily contributed by smartphone users who installed specific applications.

The dataset is organized into records, each containing information about a single tower. The pieces of information we used in this paper are the following:

- **Identifier**: each tower has associated an unique cell identifier;
- **Position**: latitude and longitude coordinates;
- **Radio technology**: cellular system supported by the tower, such as LTE, UMTS, GSM, etc.;
- **Range**: estimated range of the tower in meters;
- **Samples**: number of different observations obtained for the tower.

OpenCelliD covers Europe, most of North America, India, South East Asia, and some parts of South America and Africa. For this paper, we used cell towers in Pisa, Italy.

[1] a comprehensive list of papers that use it can be found here: https://github.com/swinedge/eua-dataset

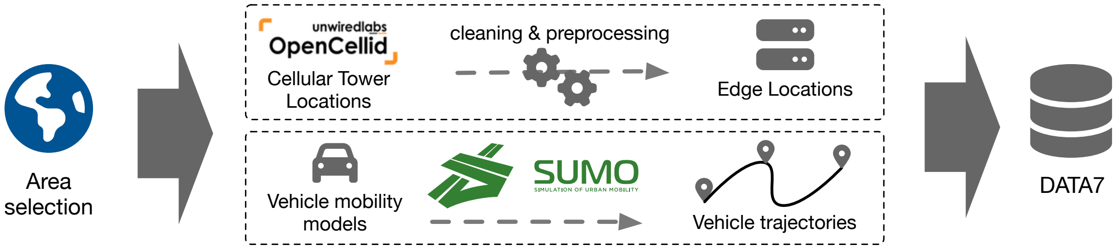
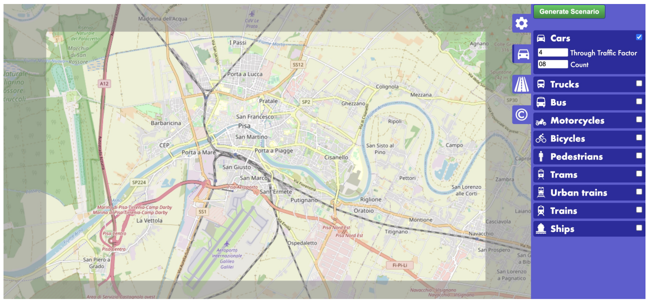


**Figure 1: Methodology for the DATA7 construction**



☑ Cars
4 Through Traffic Factor
08 Count
□ Trucks
□ Bus
□ Motorcycles
□ Bicycles
□ Pedestrians
□ Trams
□ Urban trains
□ Trains
□ Ships

**Figure 2: OSMWebWizard tool for Pisa city routes**

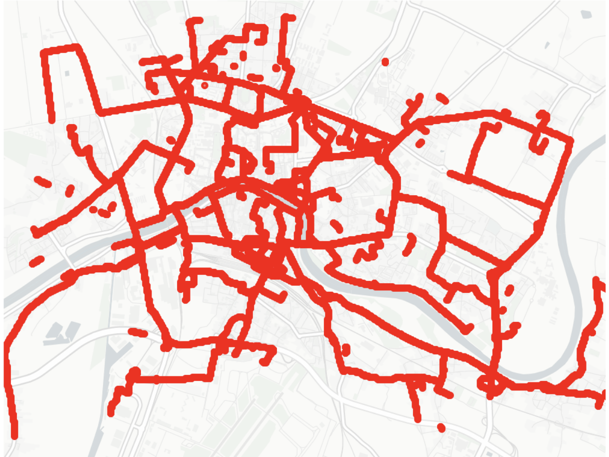


**Figure 3: Traffic**

Estimating edge devices positions required a few preprocessing steps to eliminate errors in the raw OpenCelliD database. Being OpenCelliD a community-driven project, the localization and information about towers' location, range, and characteristics can be imprecise [18].

First, several observations reported a range exceeding the maximum theoretical working range for cellular towers, and we simply discarded these observations. Second, the location of the same tower could differ among different measurements, due to location errors in smartphones and distance estimation errors. Therefore, the position of each tower is computed as the average of the observed positions.

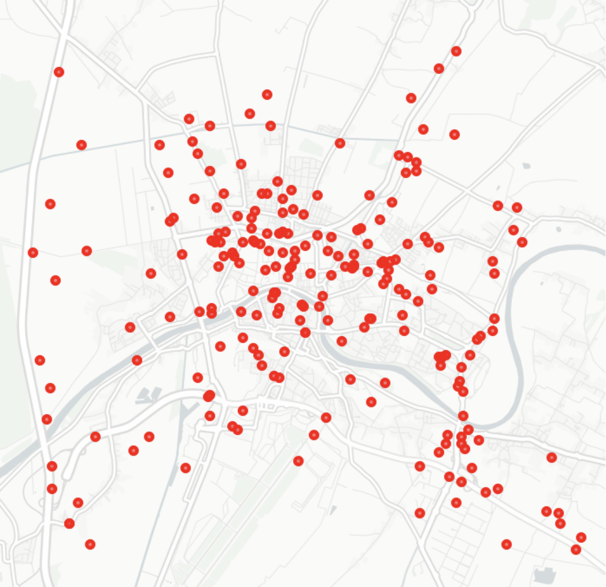


**Figure 4: Edge devices positions in the area of Pisa.**

Third, we removed those observations with very few samples (i.e. $< 3$) as they probably do not contain reliable information.

The resulting dataset contains 216 cell towers, distributed in the area of Pisa as in Figure 4.

### 3.3 DATA7 description

The dataset contains a record for each observation of a vehicle in the range of an edge device. The same vehicle can be in the range of multiple Edge nodes at the same time.

In the simulation scenario of SUMO, there were a large number of vehicles involved (3500). However, to streamline our analysis and focus on the most relevant data, we randomly selected 630 of them to be included in our dataset. The chosen vehicles are exemplary cases of edge system users actively using Edge services. However, it's crucial to recognize that their mobility behaviors are shaped by the prevailing mobility patterns of the city. This implies that despite operating under the edge devices system, these users actions still align with the city's chosen mobility model. This focused approach allows us to understand better how these specific entities interact with the edge network devices and how their use may affect them. For example, Figure 5 shows the workload of a selected edge node

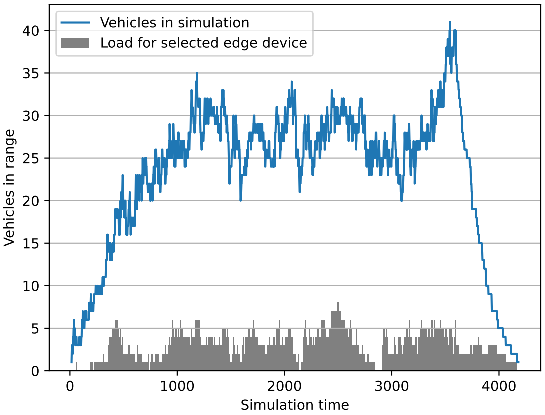


**Figure 5: Workload as the number of connected users to a selected Edge node over time (grey, bottom). Total number of vehicles in the simulation (blue, top).**

| Field | Description |
|---|---|
| edge_id | Unique identifier of the edge devices |
| edge_lat | Latitude coordinate of the edge device |
| edge_lon | Longitude coordinate of the edge device |
| time | Simulation step of the observation |
| vehicle_id | Unique identifier of the vehicle |
| vehicle_lat | Latitude coordinate of the vehicle |
| vehicle_lon | Longitude coordinate of the vehicle |
| distance | Geodesic distance in meters from the vehicle and the edge device |

**Table 1: DATA7 schema**

compared to the total number of vehicles in the simulation at any given time. The final dataset is composed of about 730k records (whose format is presented in Table 1) that amount to 60MB. No anonymization has been performed as no personal data is involved in the creation of the dataset. Finally, the dataset is publicly available as a comma-separated value (CSV) file [4].

## 4 CONCLUSION

This paper presented DATA7, a comprehensive collection of data on edge devices and mobility patterns, which aims to provide researchers and practitioners focusing on resource management at the edge to develop and evaluate solutions in a realistic environment. The dataset includes real data concerning the positioning of the antennas of mobile network carriers and simulated data obtained by leveraging realistic mobility patterns generated by SUMO.

## ACKNOWLEDGMENTS

This work has been partially supported by the ACCORDION project (EC grant n.871793) and by "Nextgen Applications" (joint project of CNR-ISTI and CS Dept of University of Pisa n.1650/2022).

## REFERENCES

[1] Abdullah Alsaedi, Nour Moustafa, Zahir Tari, Abdun Mahmood, and Adnan Anwar. 2020. TON_IoT Telemetry Dataset: A New Generation Dataset of IoT and IIoT for Data-Driven Intrusion Detection Systems. *IEEE Access* 8 (2020), 165130–165150. https://doi.org/10.1109/ACCESS.2020.3022862

[2] Jörn Altmann, Ram Govinda Aryal, Emanuele Carlini, Cheolyong Cho, Massimo Coppola, Patrizio Dazzi, Netsanet Haile, Young-Woo Jung, Burak Karaboga, Sun-Wook Kim, Myungjin Kim, Won-Bon Koo, Corrina Lechler, Lara Lopez, Iain James Marshall, Charles Lee Thoma Marshall, Kélian Marshall, Enric Pages, Evangelos Psomakelis, Ganis Zulfa Santoso, Antonia Schwichtenberg, Song-Woo Sok, Konstantinos Tserpes, Theodora Varvarigou, Ioannis Violos, Richard Wacker, and Thorsten Zylowski. 2019. *BASMATI WiFi Localization Dataset.* https://doi.org/10.5281/zenodo.3333032

[3] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. 2018. Microscopic Traffic Simulation using SUMO. In *IEEE Intelligent Transportation Systems Conference (ITSC).*

[4] Emanuele Carlini, Massimo Coppola, Patrizio Dazzi, Luca Ferrucci, Hanna Kavalionak, and Matteo Mordacchini. 2023. *DATA7: A dataset that uses synthetic trajectories of vehicles and real cellular tower locations to simulate the workload of Edge nodes in the city of Pisa.* https://doi.org/10.5281/zenodo.7806928

[5] Xianbang Diao, Wendong Yang, Lianxin Yang, and Yueming Cai. 2021. Uav-relaying-assisted multi-access edge computing with multi-antenna base station: Offloading and scheduling optimization. *IEEE Transactions on Vehicular Technology* 70, 9 (2021), 9495–9509.

[6] Eucalyptus. 2014. Eucalyptus Traces. https://sites.cs.ucsb.edu/ rich/workload/.

[7] OpenStreetMap Foundation. 2023. OpenStreetMap. https://www.openstreetmap.org/.

[8] Google. 2019. Google Traces. https://github.com/google/cluster-data.

[9] Akshay Jajoo, Y. Charlie Hu, Xiaojun Lin, and Nan Deng. 2021. The Case for Task Sampling based Learning for Cluster Job Scheduling. *Computing Research Repository* abs/2108.10464 (2021). arXiv:2108.10464 https://arxiv.org/abs/2108.10464

[10] Akshay Jajoo, Y. Charlie Hu, Xiaojun Lin, and Nan Deng. 2022. A Case for Task Sampling based Learning for Cluster Job Scheduling. In *19th USENIX Symposium on Networked Systems Design and Implementation (NSDI 22).* USENIX Association, Renton, WA, USA. https://www.usenix.org/conference/nsdi22/presentation/jajoo

[11] Wazir Zada Khan, Ejaz Ahmed, Saqib Hakak, Ibrar Yaqoob, and Arif Ahmed. 2019. Edge computing: A survey. *Future Generation Computer Systems* 97 (2019), 219–235.

[12] Daniel Krajzewicz. 2010. Traffic simulation with SUMO–simulation of urban mobility. *Fundamentals of traffic simulation* (2010), 269–293.

[13] Phu Lai, Qiang He, Mohamed Abdelrazek, Feifei Chen, John Hosking, John Grundy, and Yun Yang. 2018. Optimal Edge User Allocation in Edge Computing with Variable Sized Vector Bin Packing. In *Service-Oriented Computing*, Claus Pahl, Maja Vukovic, Jianwei Yin, and Qi Yu (Eds.). Springer International Publishing, Cham, 230–245.

[14] Matteo Mordacchini, Luca Ferrucci, Emanuele Carlini, Hanna Kavalionak, Massimo Coppola, and Patrizio Dazzi. 2021. Self-organizing Energy-Minimization Placement of QoE-Constrained Services at the Edge. In *Economics of Grids, Clouds, Systems, and Services*, Konstantinos Tserpes, Jörn Altmann, José Ángel Bañares, Orna Agmon Ben-Yehuda, Karim Djemame, Vlado Stankovski, and Bruno Tuffin (Eds.). Springer International Publishing, Cham, 133–142.

[15] Institute of Transportation Systems. 2023. osmWebWizard. https://sumo.dlr.de/docs/Tutorials/OSMWebWizard.html.

[16] Institute of Transportation Systems. 2023. SUMO-GUI. https://sumo.dlr.de/docs/sumo-gui.html.

[17] OpenCellID. 2023. The world's largest Open Database of Cell Towers. https://opencellid.org.

[18] Michael Ulm, Peter Widhalm, and Norbert Brändle. 2015. Characterization of mobile phone localization errors with OpenCellID data. In *2015 4th International Conference on Advanced Logistics and Transport (ICALT).* IEEE, 100–104.

[19] John Wilkes. 2020. *Google cluster-usage traces v3.* Technical Report. Google Inc., Mountain View, CA, USA. Posted at https://github.com/google/cluster-data/blob/master/ClusterData2019.md.

[20] John Wilkes. 2020. Yet more Google compute cluster trace data. Google research blog. Posted at https://ai.googleblog.com/2020/04/yet-more-google-compute-cluster-trace.html..

[21] Junjuan Xia, Lisheng Fan, Nan Yang, Yansha Deng, Trung Q Duong, George K Karagiannidis, and Arumugam Nallanathan. 2020. Opportunistic access point selection for mobile edge computing networks. *IEEE Transactions on Wireless Communications* 20, 1 (2020), 695–709.

[22] Bin Xiang, Jocelyne Elias, Fabio Martignon, and Elisabetta Di Nitto. 2021. A dataset for mobile edge computing network topologies. *Data in Brief* 39 (2021), 107557. https://doi.org/10.1016/j.dib.2021.107557

[23] Georgios Zachos, Ismael Essop, Georgios Mantas, Kyriakos Porfyrakis, Josè C. Ribeiro, and Jonathan Rodriguez. 2021. Generating IoT Edge Network Datasets based on the TON_IoT Telemetry Dataset. In *2021 IEEE 26th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD).* 1–6. https://doi.org/10.1109/CAMAD52502.2021.9617799