Data Article

Multivariate sensor signals collected by aquatic drones involved in water monitoring: A complete dataset

Alberto Castellini [a],*, Domenico Bloisi [b], Jason Blum [a], Francesco Masillo [a], Alessandro Farinelli [a]

[a] Department of Computer Science, University of Verona, Strada le Grazie 15, 37134 Verona, Italy
[b] Department of Mathematics, Computer Science, and Economics, University of Basilicata, Viale dell'AteneoLucano, 10, 85100 Potenza, Italy

ARTICLE INFO

ABSTRACT

Sensor data generated by intelligent systems, such as autonomous robots, smart buildings and other systems based on artificial intelligence, represent valuable sources of knowledge in today's data-driven society, since they contain information about the situations these systems face during their operation. These data are usually multivariate time series since modern technologies enable the simultaneous acquisition of multiple signals during long periods of time. In this paper we present a dataset containing sensor traces of six data acquisition campaigns performed by autonomous aquatic drones involved in water monitoring. A total of 5.6 h of navigation are available, with data coming from both lakes and rivers, and from different locations in Italy and Spain. The monitored variables concern both the internal state of the drone (e.g., battery voltage, GPS position and signals to propellers) and the state of the water (e.g., temperature, dissolved oxygen and electrical conductivity). Data were collected in the context of the EU-funded Horizon 2020 project INTCATCH (http://www.intcatch.eu) which aims to develop a new paradigm for monitoring water quality of catchments. The aquatic drones used for data acquisition are Platypus Lutra boats. Both autonomous and manual

---

* Corresponding author.
  E-mail address: alberto.castellini@univr.it (A. Castellini).

drive is used in different parts of the navigation. The dataset is analyzed in the paper "Time series segmentation for state-model generation of autonomous aquatic drones: A systematic framework" [1] by means of recent time series clustering/segmentation techniques to extract data-driven models of the situations faced by the drones in the data acquisition campaigns. These data have strong potential for reuse in other kinds of data analysis and evaluation of machine learning methods on real-world datasets [2]. Moreover, we consider this dataset valuable also for the variety of situations faced by the drone, from which machine learning techniques can learn behavioral patterns or detect anomalous activities. We also provide manual labeling for some known states of the drones, such as, drone inside/outside the water, upstream/downstream navigation, manual/autonomous drive, and drone turning, that represent a ground truth for validation purposes. Finally, the real-world nature of the dataset makes it more challenging for machine learning methods because it contains noisy samples collected while the drone was exposed to atmospheric agents and uncertain water flow conditions.

Specifications table

Subject	Artificial Intelligence
Specific subject area	Activity recognition, situation assessment, time series analysis, sensor data analysis, analysis of data collected by autonomous surface vessels
Type of data	Table
How data were acquired	Sensors installed on aquatic drones (Platypus Lutra boats). A smartphone on-board is used to provide the GPS information. A sensor control unit (Go-Sys BlueBox) is connected to an Arduino e-board and transmits information coming from the sensors measuring the water quality indicators including temperature, electrical conductivity and dissolved oxygen. The BlueBox sends geolocalized data over the Internet to a database called WAIS (Water Information System).
Data format	Raw Analyzed (re-sampling and interpolation)
Parameters for data collection	A total of 5.6 h of navigation are available, with data coming from both (artificial and natural) lakes and rivers, and from different locations in Italy and Spain. The monitored variables concern both the internal state of the drone (e.g., battery voltage, GPS position and signals to propellers) and parameters of the water (e.g., temperature, dissolved oxygen and electrical conductivity). A (post-processing) manually generated and partial labeling is also provided for four known situations, namely, drone inside/outside of the water, upstream/downstream/no stream navigation, manual/autonomous drive, and turning.
Description of data collection	The aquatic drone is put into the water and driven manually or autonomously to collect data. In case of manual navigation an RC controller is used by the operator to drive the boat. In case of autonomous navigation the operator can interact with the boat using a tablet. An Android Control app allows to define, using a Grafical User Interface, paths to be travelled by the boat. For example, it is possible to create a mission path by selecting a set of waypoints on a map with few clicks. The speed of the drone can be set via the Android app and the mission can be started, paused and stopped using the interface. During the mission, sensors acquire data at their pre-defined sampling frequency. Data are stored within a memory device in the drone or sent over the Internet to a catchment database called WAIS (Water Information System).

Data source location	- Location 1 (ESP2): River Ter, Torellò, Barcelona, Spain (latitude 42°01′45.6″N, longitude 2°15′13.1″E)
	- Location 2 (ESP5): River Ter, Manlleu, Barcelona,Spain (latitude 41°59′56.1″N, longitude 2°17′24.2″E)
	- Location 3 (ESP4): Pantà de Sau (lake), Sau reservoir, Barcelona, Spain (latitude 41°58′10.3″N, longitude 2°24′48.7″E)
	- Location 4 (GARDA3): Lake Garda, Verona, Italy (latitude 45°34′39.5″N, longitude 10°42′04.2″E)
	- Location 5 (ITA1): Atlantide fishing pond, Verona, Italy (latitude 45°21′32.4″N, longitude 11°01′14.3″E)
	- Location 6 (ITA6): Atlantide fishing pond, Verona, Italy (latitude 45°21′32.1″N, longitude 11°01′14.5″E)
Data accessibility	Repository name: Intcatch aquatic drone sensor dataset
	Direct URL to data:
	https://data.mendeley.com/datasets/gtt7stf5x8 (Mendeley dataset)
	http://profs.scienze.univr.it/~castellini/intcatchsensordataset.html (Private website repository)
Related research article	A. Castellini, M. Bicego, F. Masillo, M. Zuccotto, A. Farinelli, Time series segmentation for state-model generation of autonomous aquatic drones: A systematic framework. Engineering Applications of Artificial Intelligence (2020) 103,499, https://doi.org/10.1016/j.engappai.2020.103499

Value of the data

- The dataset can be useful to test and evaluate machine learning algorithms for multivariate time series analysis and prediction. Since data contain both drone state and water quality variables, they can be used also to test data analysis methods in both robotics and environmental monitoring contexts. Notice that water parameters were not validated by water quality experts but they were only collected for testing the overall ASV architecture, hence these data cannot be used to draw conclusions about the water quality of analyzed catchments.

- Researchers in applied disciplines related to time series data analysis can benefit from these data since they can use them to (i) investigate behaviors of real aquatic drones (e.g., battery discharge, GPS localization, motion in different environments), (ii) investigate properties of water in real catchments. Also Drone manufacturers might benefit from these data by analyzing drone equipment behaviors and identifying possible performance improvement in these equipments (e.g., propellers, batteries).

- The analysis of this dataset can suggest the development of new experiments for improving autonomous water monitoring.

- The additional value of this dataset is the simultaneous presence of signals from drone equipments and water properties.

1. Data description

Autonomous surface vehicles (ASVs) have been recently introduced in persistent large scale water monitoring of aquatic environments because of their capability to support in an efficient way traditional manual sampling approaches [3,4]. The dataset here presented is a collection of six multiple-sensor traces generated in independent data acquisition campaigns [6], also called experiments in the following. We provide one *comma separated value (csv)* file for each campaign. Table 1 summarizes the main properties of each experiment, namely, the sequential *id* we assigned to the campaign, the *campaign name*, the *location* where the campaign is performed, the *type of catchment* (i.e., river or lake) in which the campaign is performed, the start and end *date/time*, the *duration* (in minutes) of the campaign and the *number of samples* available for the campaign. The overall dataset contains 20,187 samples for a total of 5.6 h of navigation. The sampling frequency is 1 Hz. The variables (i.e., columns in the *csv* files) available for each campaign are eleven and they are described in Table 2 [1]. We notice that the signals (*m0_current* and

Table 1
Summary of data campaigns.

Id	Campaign name	Campaign location	Lake/ River	Campaign date/time	Duration	Samples
1	ESP2	River Ter, Torellò, Barcelona, Spain	R	2017-03-31, 10:15:13 2017-03-31, 11:02:06	47′	2814
2	ESP5	River Ter, Manlleu, Barcelona, Spain	R	2017-03-31, 07:28:33 2017-03-31, 08:28:33	60′	3601
3	ESP4	Pantà de Sau, Sau reservoir, Barcelona, Spain	L	2017-03-30, 08:22:25 2017-03-30 09:01:58	39′	2374
4	GARDA3	Lake Garda, Verona, Italy	L	2017-05-09, 09:43:17 2017-05-09, 10:24:07	40′	2451
5	ITA1	Atlantide fishing pond, Verona, Italy	L	2017-04-20, 09:15:21 2017-04-20, 11:16:03	121′	7243
6	ITA6	Atlantide fishing pond, Verona, Italy	L	2017-03-07, 09:52:49 2017-03-07, 10:21:12	28′	1704

Table 2
Variable descriptions.

Id	Variable name	Description	Unit	Example
1	*date_time*	Date and time in which the sample is acquired	–	2017-03-31 10:15:13
2	*latitude*	Latitude of the location where the sample is acquired	°	42.0293361082
3	*longitude*	Longitude of the location where the sample is acquired	°	2.25363680341
4	*altitude*	Height above sea level	m	81.9
5	*ec*	Water electrical conductivity	μS/cm	518.333333333
6	*temp*	Water temperature	°C	13.1666666667
7	*do*	Water dissolved oxygen	mg/l	10.4266666667
8	*voltage*	Drone's battery voltage	V	16.915
9	*m0_current*	Signal to propeller 0	–	0.8913
10	*m1_current*	Signal to propeller 1	–	0.8913
11	*heading*	Compass direction in which the drone's bow is pointed	°	89.781208369

m1_current) sent by the controller to the two propellers, follow a differential-based schema, and they can have values between −1 and +1, where −1 means that the engine provides maximum backward power and +1 means that it provides maximum forward power. This mechanism is used to make the drone turn (when the two signals have different intensity) or going straight (when the two signals have similar intensity), as explained in the next section. In addition to sensor data, we provide for each campaign also four labelings about basic states of the drone, namely, drone inside or outside of the water (*water*), drone navigating upstream, downstream or with no water flow (*flow*), drone manually or autonomously driven (*drive*), and drone turning (*curve*). These labelings were manually generated by observing the drone paths in geographical maps during a post-processing phase and merging this information with notes taken during the experiments. For each experiment we provide a *csv* file for each labeling (i.e., four csv labeling files for each experiment). Since this manual labeling process is very complex and error-prone, we labelled only the parts of the paths where secure information was available and used label '*0*′ to identify unlabelled samples, hence many labelings are partial.

In the repository, data files are organized by acquisition campaigns (i.e., experiments). Once a specific campaign is selected (e.g., ESP2), the user finds two folders called *Labels* and *SensorData* respectively. Folder *SensorData* contains a *csv* file called *ExpName_sensorData.csv* (where *ExpName* is the experiment name, e.g., ESP2) providing the sensor data matrix for the specific experiment (see Tables 1 and 2 for details about the number of samples and the variables in each matrix). Folder *Labels* contains four subfolders, one for each labeling, called respectively *ExpName_curves, ExpName_drive, ExpName_flow* and *ExpName_water*. Each of these folders contains a *csv* file called *ExpName_LabelingName_labels.csv* providing the labeling vector *LabelingName* (e.g., curve) for experiment *ExpName* (e.g., ESP2), and a *txt* file called *ESP2_curves_label_legend.txt* providing the legend for this labeling. The folder structure described so far is replicated for each experiment
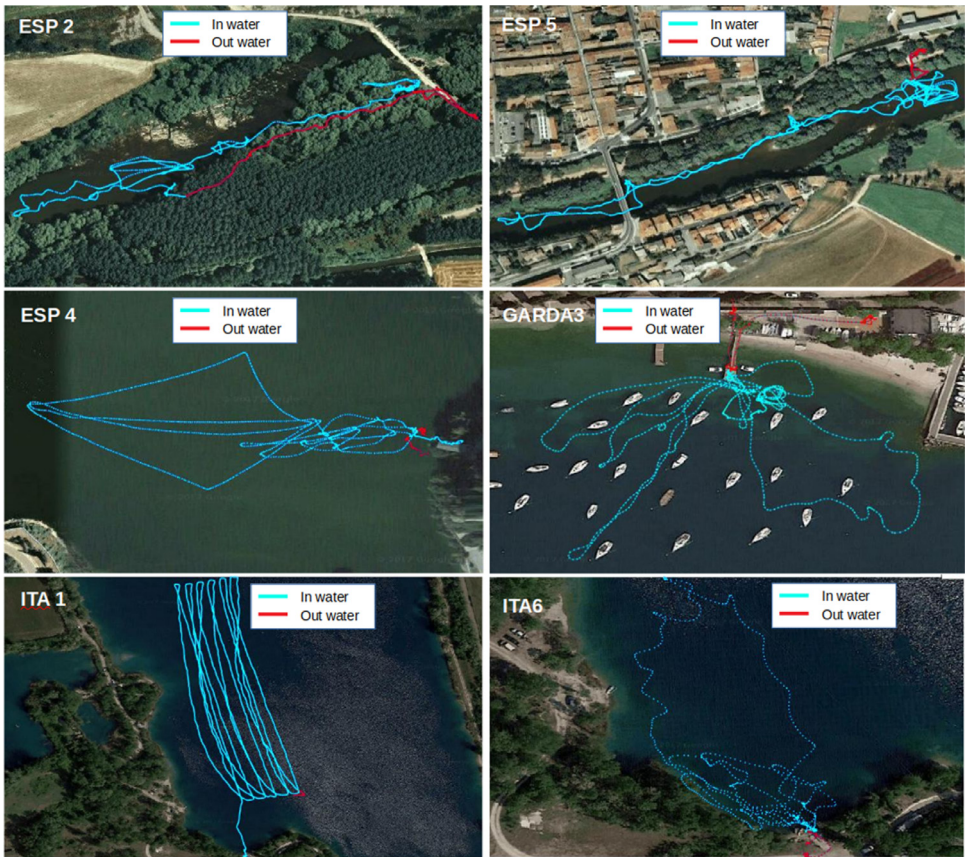
Fig. 1. Geolocalization and shape of the data acquisition paths. Colors represent labeling *"drone inside/outside of the water"*.

and for each labeling hence the user can easily recover the sensor data files and the labeling files that it needs for his/her analysis.

As a descriptive analysis of the dataset, we display two figures, one showing the geolocalization and the shape in a map of the path for each campaign (see Fig. 1) and another showing the main statistics of water parameters and drone parameters for each campaign (see Fig. 2). Fig. 1 displays also, as an example, the labeling *"drone inside/outside of the water"* by means of the path color, where red indicates samples taken outside of the water (for instance, before or after the beginning of the campaign) and light blue indicates samples taken inside the water. The figure provides useful understanding about the shape of the path and the kind of environment in which the acquisition was performed [6]. Fig. 2 instead provides boxplots for water parameters (i.e., *ec, temp, do*) and drone parameters (i.e., *voltage, m0_current, m1_current, heading*) in different experiments [5]. The purpose of this figure is to provide the reader with an overview of the parameters related to the drone and to the water, with specific focus on how they vary across the different experiments. These charts show, in an intuitive way, the median (red horizontal lines), the lower (Q1) and upper (Q3) quartiles (upper and lower extremes of the boxes), the interquartile range (IQR=Q3-Q1), the range of the data (whiskers in the bottom: Q1 – 1.5*IQR and whiskers on top: Q3 + 1.5*IQR), outliers (defined as points out of the whiskers), minimum and maximum values, for each variable and for each experiment, hence they provide valuable information about the dataset in a very concise way.
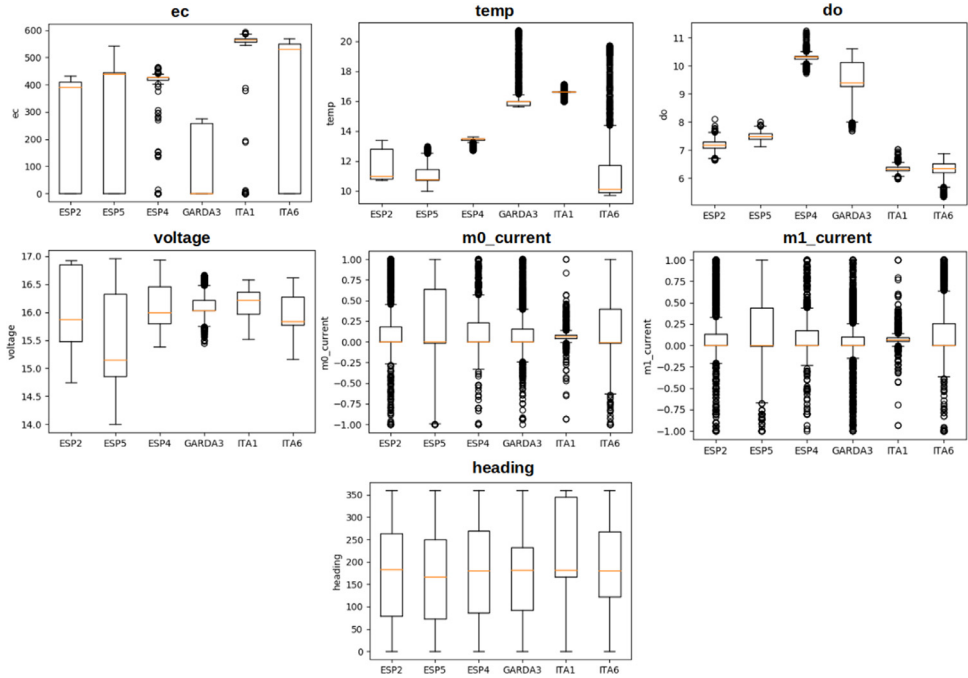
Fig. 2. Boxplots of water parameters (i.e., ec, temp, do) and drone parameters (voltage, m0_current, m1_current, heading) for each campaign. They provide basic descriptive statistics about the main variables in the dataset.

2. Experimental design, materials, and methods

Data acquisition campaigns are performed by Lutra mono hull drones produced by Platypus [8] and engineered in the EU Horizon 2020 INTCATCH project [9] to perform water monitoring of catchments. The boats are about 1 m long and 0.5 m wide. The main equipments of the drone and the tools used to collect data are represented in Fig. 3 and described in the following.

Lutra boats used in these experiments have in-water *propellers* hence they need a water depth of at least 25 cm to be deployed. The drone motion is based on a differential drive mechanism, therefore the drone moves straight forward when the signals to propellers have the same positive values, it moves straight backward when the signals to propellers have the same negative values, and it turns when the signals to propellers have values of different sign. The drone is equipped with a four cells lithium polymer (LiPo) *battery* with capacity of 16,000 mAh and a nominal voltage of 14.8 V (10C discharge). The experiments were performed using two types of interaction, namely, *manual drive* in which the operator interacts with the drone by an RC controller, and *autonomous drive* which requires the operator to pre-define a path by setting waypoints in a map on a tablet (by means of a specific Android Control app) and then to start/stop the navigation by sending specific commands. During autonomous navigation the desired trajectory is maintained by constantly comparing the real location measured by the GPS with the desired position (defined as the shortest distance to the desired path). A *controller* adjusts the commands to propellers to correct the route in case of divergence. The *localization* and orientation of the drone is measured by a smartphone positioned into the boat which provides GPS, compass and gyroscope information. *Sensor* management and sensor data transmission to the cloud are performed by a Go-Sys BlueBox [10] gateway unit connected to an Arduino e-board. Two sensors placed under the hull allow to measure three water parameters, namely,
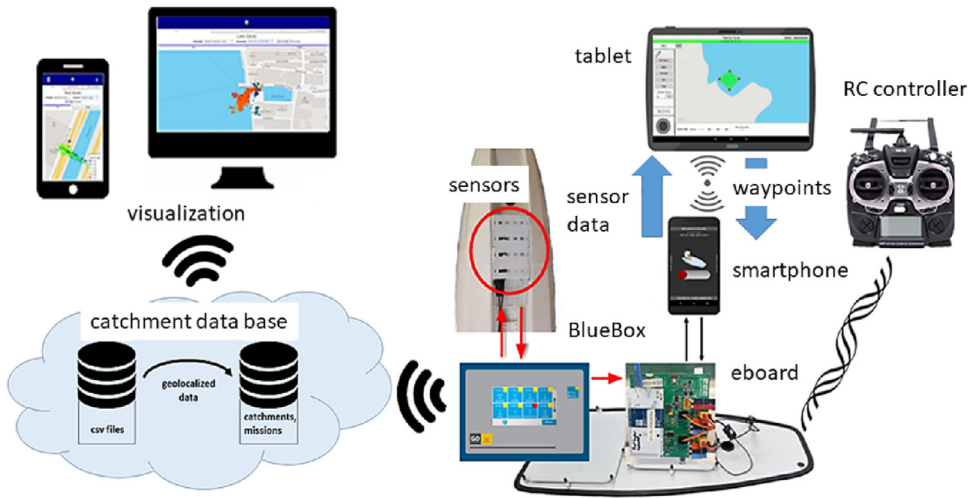
Fig. 3. Drone architecture: main tools and equipments [7].

electrical conductivity, temperature and dissolved oxygen (temperature was measured with the same sensor with which the conductivity was measured). The sensors used to measure water electrical conductivity and dissolved oxygen are A*tlas Scientific tools,* whose technical properties can be found in [11]. Geolocalized sensor signals are sent by the Go-Sys BlueBox to a *database* called WAIS (Water Information System) over the Internet. From these database log files can be retrieved containing the sensor traces of interest. *Log-files* are automatically generated using proprietary libraries that generate a matrix of time series, in *csv* format, having one sensor signal for each column and time instants in rows. Since different sensors may have different sampling frequencies a proprietary software is used to extract interpolated and re-sampled data with 1 Hz sampling frequency.

Experiments are performed by bringing the boat near to the catchment of interest, turning on the equipment before deploying the drone into the water (for these reasons the dataset contains samples taken out of the water, which are identifiable by very low electrical conductivity), and then deploying the boat into the water after testing the equipments. As a final remark, we notice that water parameters were not validated by water quality experts but they were only collected for testing the overall ASV architecture, hence these data cannot be used to draw conclusions about the water quality of analyzed catchments.

Acknowledgments

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.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2020.105436.

References

[1] A. Castellini, M. Bicego, F. Masillo, M. Zuccotto, A. Farinelli, Time series segmentation for state-model generation of autonomous aquatic drones: a systematic framework, Eng. Appl. Artif. Intell. 90 (2020) 103499, doi:10.1016/j.engappai.2020.103499.
[2] European Policy Evaluation Consortium (E.P.E.C), Detailed Assessment of the Market Potential, and Demand for, an EU ETV Scheme (2013) https://publications.europa.eu/s/mNoC.
[3] M. Dunbabin, A. Grinham, Quantifying spatiotemporal greenhouse gasemissions using autonomous surface vehicles, J. Field Robot. 1 (2017) 151–169, doi:10.1002/rob.21665.
[4] D.L. Codiga, A marine autonomous surface craft for long-duration, spatially explicit, multidisciplinary water column sampling in coastal and estuarine systems, J. Atmos. Ocean. Technol. 3 (2015) 627–641.
[5] A. Castellini, M. Bicego, D. Bloisi, J. Blum, F. Masillo, S. Peignier, A. Farinelli, Subspace Clustering For Situation Assessment in Aquatic drones: a Sensitivity Analysis For State-Model improvement, Cybernetics and Systems, 50, Taylor & Francis, 2019, pp. 658–671, doi:10.1080/01969722.2019.1677333.
[6] A. Castellini, F. Masillo, M. Bicego, D. Bloisi, J. Blum, A. Farinelli, S. Peigner, Subspace clustering for situation assessment in aquatic drones, in: Proceedings of the Symposium on Applied Computing, SAC 2019, 2019, ACM, 2019, pp. 930–937, doi:10.1145/3297280.3297372.
[7] L. Steccanella, D.D. Bloisi, A. Castellini, A. Farinelli, Waterline and obstacle detection in images from low-cost autonomous boats for environmental monitoring, Robotics and Autonomous Systems, 124, Elsevier, 2020, doi:10.1016/j.robot.2019.103346.
[8] Platypus, LLC website. http://senseplatypus.com.
[9] EU Horizon 2020 INTCATCH Project Website. http://www.intcatch.eu.
[10] Go-sys Bluebox website: https://www.go-sys.de/en/bluebox/.
[11] Atlas scientific sensor website: https://www.atlas-scientific.com.