IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received June 5, 2021, accepted June 15, 2021, date of publication June 18, 2021, date of current version June 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3090582

Human Mobility Prediction With Region-Based Flows and Water Consumption

FERNANDO TERROSO-SÁENZ[1], ANDRÉS MUÑOZ[1], JULIO FERNÁNDEZ-PEDAUYE[2], AND JOSÉ M. CECILIA[2]

[1]Polytechnical School, Universidad Católica de Murcia, 31107 Murcia, Spain
[2]Department of Computer Engineering, Universitat Politècnica de València, 46022 Valencia, Spain

Corresponding author: Fernando Terroso-Sáenz (fterroso@ucam.edu)

This work was supported in part by the Spanish Ministry of Science and Innovation, through the Ramon y Cajal Program under Grant RYC2018-025580-I, Grant RTI2018-096384-B-I00, and Grant RTC-2017-6389-5; in part by the Fundación Séneca del Centro de Coordinación de la Investigación de la Región de Murcia under Project 20813/PI/18, and in part by the ''Conselleria de Educación, Investigación, Cultura y Deporte, Direcció General de Ciéncia i Investigació, Proyectos AICO/2020,'' Spain, under Grant AICO/2020/302.

ABSTRACT We are witnessing an increasing need to accurately measure people's mobility as it has become an instrumental factor for the development of innovative services in multiple domains. In this context, several ICT solutions have relied on location-based technologies such as GPS, WiFi or Bluetooth to track individual's movements. However, these technologies are limited by the privacy restrictions of data providers. In this paper we propose a methodology to robustly predict citizens' mobility patterns based on heterogeneous data from different sources. Particularly, our methodology focuses on a human mobility predictor based on a low-resolution mobility dataset and the use of water consumption data as a facilitator of this prediction task. As a result, this work explores whether the water consumption within a geographical region can reveal human activity patterns relevant from the point of view of the mobility mining discipline. This approach has been tested in a residential area near Madrid (Spain) obtaining quite promising results.

INDEX TERMS Human mobility, water consumption, location data, forecasting methods.

I. INTRODUCTION

Nowadays, the Internet of Things (IoT) has completely transformed modern societies. One clear effect of this impact is the fact that most of the regular objects and artifacts that we use and wear every day, from bracelets to cars, are now equipped with location-enabling technologies like GPS, WiFi or Bluetooth able to position such objects in real-world physical places.

In this context, the *human-mobility mining* discipline has emerged as one of the most important research trends in the vast data science and artificial intelligence ecosystem [1]. Basically, this field seeks for extracting meaningful knowledge about human movement behaviours at different temporal and spatial scales. One of the most relevant findings in this discipline is that human mobility is quite predictable at some extend [2]. As a result, the prediction of where and when people is going to move is an instrumental tool in domains like healthcare [3], urban services [4] and transportation management [5].

One important factor when it comes to develop such forecasting methods is the location data that is going to feed the system. Thus, it is possible to find related proposals based on GPS traces [6], Call Detail Records (CDRs) [7] or Online Social Media (OSM) posts [8] covering different temporal and spatial scenarios.

Nevertheless, it is possible to observe two important limitations in many existing methods for human-mobility forecasting,

- On the one hand, these solutions basically rely on the raw spatio-temporal trajectories generated by different moving *objects* like taxis or individuals reporting their current location every few seconds or minutes. Nonetheless, in a real-world scenario this type of high-quality location data is, in most occasions, rather inaccessible due to several privacy and economic policies defined by data providers and operators [9]. At the same time, the open data movement has promoted the release of an increasing number of human-mobility datasets [10]–[12]. However, such open availability comes at the cost of data filtering and aggregation stages before the dataset is being published to complain with several

The associate editor coordinating the review of this manuscript and approving it for publication was Muhammad Sharif.

restrictions. This is because location data is quite sensible in terms of privacy. The development of predictors leveraging such coarse-grained mobility data is still scarce in the mobility mining domain.

- On the other hand, the mobility behaviour of a population is strongly related to their latent activities at each moment [13], [14]. However, existing solutions for human-mobility prediction usually neglect the usage of human-activity data which is not directly related to movement or displacement actions, that is, the spatio-temporal traces generated by the target moving objects. Indeed, most works combine different types of mobility feeds (e.g. taxis, buses, bikes) [15], [16] or contextual urban data like Points of Interest (POIs) [17] for this purpose. However, the combination of heterogeneous data sources, not directly coming from the mobility domain, could help to obtain more robust predictors.

Taking into account the aforementioned limitations, the present work proposes a novel mechanism to predict the human mobility behaviour in an urban area that makes use of two different types of human activity data.

Firstly, it is considered the human displacements related to the target spatial area. Instead of using high-resolution mobility data, the present work relies on an open region-based mobility feed that defines human flows at a quite large spatial scale. By using human movement data in an aggregated manner, it would be possible to deploy the solution in regions that do not have an IoT infrastructure able to capture human trips with great detail. Moreover, as a side effect of the previous reason, it can be regarded as a cost-effective mechanism. Finally, the anonymisation of the data due to the aforementioned aggregation will also relieve the privacy concerns among end-users.

Secondly, this proposal also incorporates as input data the water consumption from a set of smart meters in a large residential area allocated within the target region. The rationale of using this second dataset it that household water-consumption behaviour is an indicator of human presence and, thus, of future displacements. For example, high water-demand levels in a region at a particular moment might indicate the presence of a large number of people at home in that region. In turn, when this particular high demand of water starts decreasing, it might indicate that a large number of trips in that region is about to occur as people are leaving home after this event (e.g. after the morning shower to go to work).

By means of this approach, our work focuses on an important research gap in the mobility prediction discipline when it comes to leverage the activity patterns of a population reflected in its water consumption profile to better anticipate its future displacements. Moreover, in our setting, the human mobility data and the *exogenous* water consumption data are defined at different spatial scales. The human displacements are defined at a region level in an aggregated form whereas the water consumption data are defined at a much finer granularity as it is extracted directly from smart meters. Although water consumption and human-flow data have already been explored together in several prediction problems [18], [19], it is worth mentioning that their goal was the prediction of the water consumption of a population instead of its movement activity as in our proposal.

To do so, a 3-step methodology was adopted in this work. First, multiple time series representing different types of human flows co-occurring in the spatial region of interest were extracted from the open dataset. Then, a correlation study between the aggregated water consumption time series and such flows was performed. This allowed us to extract a target human flow. Lastly, a palette of Recurrent Neural Networks (RNNs) were fed with the selected flow and enriched with the water-consumption series. As a result, it was possible to assess the impact of this series on the prediction accuracy of the RNN models.

All in all, the goal of this paper is focused on combining heterogeneous sources using a RNN to obtain a forecasting method most robust than those obtained by just relying on mobility feeds. More in detail, the main contributions of the present work are twofold: 1) a novel human-mobility predictor based on a low-resolution mobility dataset and 2) the usage of water-consumption data as an enabler for such a forecasting task.

Finally, the remainder of the paper is structured as follows. Section 2 reviews the existing trends for human-mobility prediction and the usage of water-consumption data to uncover human activities. Then, section 3 describes the use-case setting where our solution has been deployed. In section 4, the RNN predictor is described and evaluated. Lastly, section 5 summarizes the main conclusions and potential future research lines motivated by this work.

II. RELATED WORK

The prediction of human mobility has been studied in the last years from different approaches (see [1] for a recent survey on the topic). Most of these works rely on the application of standard models and statistical techniques such as ARIMA [20], [21], linear regression models [22] or Markov models [23]. However, all these traditional methods cannot work with abnormal mobility situations such as extreme weather events or unexpected traffic incidents.

Some studies use Online Social Networks (OSN) as an additional source of data in order to discover human mobility patterns in a more dynamic manner. For example, in [24] a Twitter dataset including 10,000 unique users in New York is utilized with the aim of calculating some mobility characteristics such as distance-based displacements in the city. The results were compared with data from surveys and census of the transportation council, showing a notable similarity in some city boroughs. Similarly for the same city, Pourebrahim *et al.* [25] combined Twitter data with census data and employment statistics to predict the commuting trip distribution in New York, obtaining a more accurate model than only using the static data. Another example of using Twitter for predicting mobility patterns, in this case for detecting road-traffic events, is shown by Alomari *et al.* [26].

More recently, Deep Learning algorithms have been used for predicting human flows. In particular, Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) networks [27]–[29] are used to combine different data sources such as GPS, mobility feeds and online social networks to predict the next area in a city to be visited by tourists, the demand of public means of transport such as taxis, etc. Another type of RNN employed in this task is the Gated Recurrent Unit (GRU) model, as for example in Fan *et al.* [30]. In this work, an ensemble of GRU models, each one focusing on a particular target day, is composed in order to detect both regular and abnormal citywide mobility. Finally, the Graph Neural Networks (GNNs) are another widely used alternative for human flow forecasting. For example, in [31] a GNN is applied to predict traffic conditions based the dependencies among the road networks, whereas a similar approach is followed in [32] but using GPS trajectories and loop detectors as input data.

Regarding works related to water consumption data and human mobility, Smolak *et al.* [18], [19] compares several ML and statistical models to predict the water usage based on human mobility data. The main aim of this work is the water demand forecasting of a particular area, and they actually concluded that the use of this mobility data is correlated to the water demand and this it benefits its prediction. Di Mauro *et al.* [33] perform a review of the state-of-the-art urban water demand datasets. They reviewed 92 water demand datasets that are classified and analyzed according to the following criteria: spatial scale, temporal scale, and dataset accessibility. Barbosa *et al.* [34] reviewed several recent developments regarding mobility patterns as a collection of technical methods applicable to specific mobility-related problems.

III. SETTING OVERVIEW

The feasibility of our solution has been evaluated in a wide geographical area comprising three large *dormitory towns* near Madrid (Spain), namely, Pozuelo de Alarcón, Boadilla del Monte and Majadahonda. Fig. 1 depicts the boundaries of this area whose bounding box is defined by the latitude-longitude coordinates $\langle(40.50, -3.99), (40.35, -3.76)\rangle$.

For this spatial area it is extracted the two target datasets, namely the human-mobility open dataset and the water consumption data for each residential area. Both of them cover the same 4-month period from 18/06/2020 to 31/10/2020.

A. WATER CONSUMPTION DATASET

The water consumption dataset is obtained from smart water-meters developed by the high-tech company HIDROCONTA.[1] These meters generate anonymous water consumption data (in m^3) every 10 minutes. Specifically for this work, the data is collected from more than 200 water meters as shown in Figure 2. The raw data obtained from the water meters are curated to construct an homogeneous time series of water consumption with an hourly granularity. The data cleaning is performed as follows: First, it is identified the missing data through a histogram.Missing data are eliminated where possible, otherwise an imputation procedure is carried out. This imputation procedure replaces missing values by previous values. This is because the raw data indicates the accumulated consumption of water. If there is a period with no data from a meter, it is understood that no water has been consumed in that period, and therefore the current value amounts to the last one recorded. Finally, outliers are identified through a box plot. These outlier values are deleted and imputed as previously described.

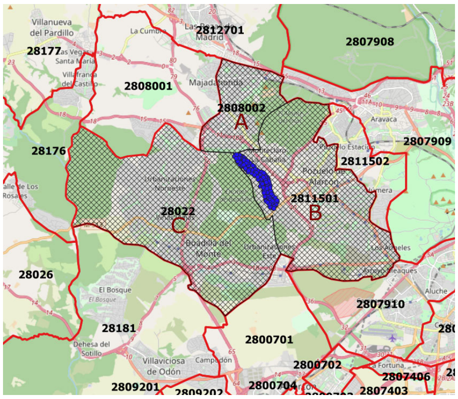


FIGURE 1. Geographical area under study. The three lined polygons (*A, B* and *C*) are the Mobility Areas (MA) under consideration whereas the red lines define the boundaries of the surrounding ones. The blue dots indicate the locations of the water meters under consideration. The numerical codes are the unique identifiers of each MA according to the Spanish Ministry of Transportation mobility survey.

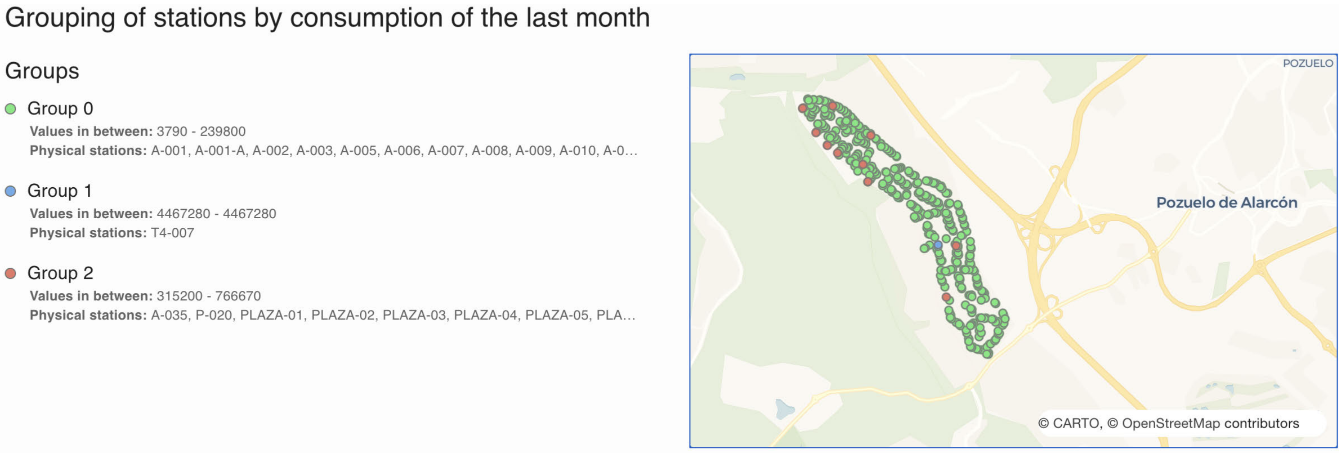
It is worth mentioning that these water meters monitor different types of installations (e.g., buildings, gardens, swimming pools, etc.). For privacy reasons, we only have access to the consumption data and no metadata about these types of installation are available. In order to overcome this problem, a clustering of the consumption data has been developed to identify those meters that have a very high consumption, which will normally be associated with garden irrigation, swimming pools, etc.; a medium consumption, which will be associated with occupied dwellings; and a low consumption, which could be associated with unoccupied dwellings, fire hydrants, etc. After this data cleaning and cleaning procedure, a water consumption time-series $\mathcal{W} = \langle w^1, w^2, .., w^{H-1}, w_1^H \rangle$ is obtained where w^i is the overall water consumption at the i-th hour from all the meters and H is the total number of hours of the dataset.

B. HUMAN MOBILITY DATASET

This dataset has been retrieved from the nation-wide human mobility dataset released by the Spanish Ministry of Transportation (SMT) in December 2020.[2] It covers a 9-month period from February 29th to November 30th, 2020 and it

[1] https://hidroconta.com/

[2] https://www.mitma.es/ministerio/covid-19/evolucion-movilidad-big-data/opendata-movilidad



Grouping of stations by consumption of the last month

Groups

● Group 0
Values in between: 3790 - 239800
Physical stations: A-001, A-001-A, A-002, A-003, A-005, A-006, A-007, A-008, A-009, A-010, A-0...

● Group 1
Values in between: 4467280 - 4467280
Physical stations: T4-007

● Group 2
Values in between: 315200 - 766670
Physical stations: A-035, P-020, PLAZA-01, PLAZA-02, PLAZA-03, PLAZA-04, PLAZA-05, PLA...

FIGURE 2. **Water meters in the geographical area. A K-means clustering is performed to group the water meters into 3 groups: High, medium and low consumption groups.**

indicates the number of trips among 3216 ad-hoc administrative areas (hereby *Mobility Areas, MA*) per hour in Spain both in its peninsular and insular extension. A *single trip* stands for the spatial displacement of an individual with distance above 500 meters. Consequently, this dataset can be regarded as a set of tuples where each one takes the form,

$$\langle date, hour, m_{origin}, m_{dest}, type_{origin}, type_{dest}, n_{trp}, dist_{range} \rangle$$

reporting that there was n_{trp} human trips from the MA m_{origin} to m_{dest} during the indicated *date* and *hour* whose covered distance in km was within the range defined by $dist_{range}$. Furthermore, the dataset also includes the *type* of origin and destination of the trips ($type_{origin}, type_{dest}$). In that sense, the survey distinguisines among *home*, *work* and *others* as possible values.

According to the official documents [35], these mobility data have been collected through Call Detail Records (CDRs) from 13 million users of an unspecified mobile-phone carrier. Once anonymised, this dataset was used to infer representative mobility statistics at the nation-level of the population of Spain and made publicly available open data. In its raw form, the dataset comprises 830,450,300 trips among MAs.

Given this dataset, we first need to filter the trips related to the target geographical area. To do so, we focused on the 3 MAs (out of the 3216 of the study) that were spatially closer to the set of water meters described in sec. III-A. Fig. 1 shows this set of three MAs, $\mathcal{M} = \langle m_A, m_B, m_C \rangle$. The land areas of these regions are 10, 34 and 47 km^2 respectively. Besides, their population is 45,165 people for m_A, 43,823 for m_B and 36,872 for m_C.

From this dataset, 21 different human flows were extracted for the MAs in $\mathcal{M}$ reflecting different human-mobility behaviours. Each flow captures a different type of human displacement for a particular combination of origin, destination and covered distance. This extraction was done by filtering the tuples of the dataset with the selection criteria defined in Table 1. Note that the concept of *inner trips* included in this table refers to those trips that occur within a particular MA, that is, whose initial and end location belong to the same area.

More in detail, each flow is structured as time series with the number of trips per hour accomplishing the flow criterion. For example, according to Table 1, flow $\mathcal{F}_1$ is defined as $\langle f_1^1, f_1^2, .., f_1^{H-1}, f_1^H \rangle$ where f_1^i is the number of human trips that arrive to m_A from any other MA to go home at the i-th hour. Likewise, $\mathcal{F}_{14} = \langle f_{14}^1, f_{14}^2, .., f_{14}^{H-1}, f_{14}^H \rangle$ where f_{14}^i is the number of trips departing from m_B to go to any other MA as long as the covered distance was less than 50 km at the i-th hour. For the sake of completeness, Appendix A comprises the time series of all the generated flows.

C. CORRELATION STUDY

Once we defined the human mobility flows indicated in Table 1, the next step was to study whether the water consumption time-series $\mathcal{W}$ described in sec. III-A were a suitable input to develop a forecasting method for any of such flows. To do so, a correlation study among these sources was performed by means of two different metrics, the Pearson's correlation coefficient (PCC) and the Mutual Information Score (MIS).

In short, PCC is a number between -1 and 1 that describes a negative or positive linear correlation, respectively. A value of zero indicates no linear correlation.

One limitation of the PCC is that is just captures the linear correlation between the variables. For that reason, the MIS was also used as it allows to measure other types of non-linear correlations [36]. In brief, MIS is a non-negative score where higher values mean higher dependency. The MIS between two variables X and Y is defined as,

$$MIS(X, Y) = H(X) + H(Y) - H(X, Y)$$

where H stands for the entropy, that is, the expected amount of information held in a variable.

The first two columns of Table 2 shows the values of these scores for each human mobility flow $\mathcal{F}$ and water

TABLE 1. Criteria followed to extract the mobility behaviours and human flows from the SMT mobility survey. The * symbol stands for *any value*. The rightmost column shows the average number of trips per hour of each flow and its standard deviation. The row in grey indicates the human flow eventually used as prediction target.

Mobility Behaviour	Flow	m_{origin}	m_{dest}	$type_{origin}$	$type_{dest}$	$dist_{range}$	Avg. num. trips
Trips to home	$\mathcal{F}_1$	*	m_A	*	home	*	134.197 (±85.494)
	$\mathcal{F}_{2/target}$	*	m_B	*	home	*	1743.095 (872.630)
	$\mathcal{F}_3$	*	m_C	*	home	*	1748.646 (±960.104)
Trips from home	$\mathcal{F}_4$	m_A	*	home	*	*	130.009 (±84.405)
	$\mathcal{F}_5$	m_B	*	home	*	*	1669.329 (±780.657)
	$\mathcal{F}_6$	m_C	*	home	*		1669.885 (±841.410)
Incoming trips	$\mathcal{F}_7$	*	m_A	*	*	*	474.741 (±256.797)
	$\mathcal{F}_8$	*	m_B	*	*	*	5312.190 (±2509.464)
	$\mathcal{F}_9$	*	m_C	*	*	*	5684.507 (±3113.019)
Outgoing trips	$\mathcal{F}_{10}$	m_A	*	*	*	*	469.887 (±244.195)
	$\mathcal{F}_{11}$	m_B	*	*	*	*	5304.106 (±2740.424)
	$\mathcal{F}_{12}$	m_C	*	*	*	*	**5687.935 (± 3346.144)**
Outgoing short trips	$\mathcal{F}_{13}$	m_A	*	*	*	<50km	828.829 (± 418.977)
	$\mathcal{F}_{14}$	m_B	*	*	*	<50km	3534.488 (±2027.222)
	$\mathcal{F}_{15}$	m_C	*	*	*	<50km	4365.407 (±2546.512)
Incoming short trips	$\mathcal{F}_{16}$	*	m_A	*	*	<50km	853.503 (±453.889)
	$\mathcal{F}_{17}$	*	m_B	*	*	<50km	3529.580 (±1891.264)
	$\mathcal{F}_{18}$	*	m_C	*	*	<50km	4362.108 (±2395.823)
Inner trips	$\mathcal{F}_{19}$	m_A	m_A	*	*	*	1005.535 (±537.541)
	$\mathcal{F}_{20}$	m_B	m_B	*	*	*	1590.357 (±714.299)
	$\mathcal{F}_{21}$	m_C	m_C	*	*	*	2531.972 (±1444.485)

consumption series $\mathcal{W}$. It can be observed that all the flows have a negative correlation with the water consumption series. Furthermore, the *trips-to-home* flows ($\mathcal{F}_1$, $\mathcal{F}_2$ and $\mathcal{F}_3$) have a slightly higher scores than the other flows. The higher PCC is obtained by flow $\mathcal{F}_{19}$ with the inner trips of the mobility area m_A.

The aforementioned correlation peak in the *incoming home* trips is meaningful as the water meters are installed in a residential area. Consequently, the home activity detected by these meters could be quite related to the movement of people going home at each moment. Likewise, the highest PCC obtained for $\mathcal{F}_{19}$ might be due to the fact that this MA has the smallest land area among the ones in $\mathcal{M}$. Consequently, the intra-MA trips contained by this flow cover quite short distances, and therefore they are probably composed by regular displacements that, again, are quite correlated with the human presence at home.

Considering both the PCC and MIS scores described before, we eventually selected a specific flow to develop a forecasting mechanism. It can be observed that, on average, $\mathcal{F}_2$ is the most correlated flow with $\mathcal{W}$ as it has a quite similar PCC score (-0.372) than the highest one (-0.380) along with the highest MIS (0.139) (see Table 2). This flow contains the incoming trips of m_A whose final destination was the travellers' home. We should remark that this flow does not only comprise commuting trips as the type of origin remains unspecified according to the selection criterion of Table 1. Hence, trips whose origin was not labelled as *work* were also considered. Furthermore, $\mathcal{F}_2$ comprises a higher number of trips per hour (1,743.095) than other flows with similar PCC like $\mathcal{F}_{16}$ (853.503) or $\mathcal{F}_{19}$ (1,005.535) (see column *avg. num. trips* in Table 1). Therefore, in operational terms, it seems a much more challenging and useful mobility flow to detect than other flows of the study. For the sake of completeness, Fig. 3 shows the time series of $\mathcal{F}_2$ (hereby $\mathcal{F}_{target}$) and $\mathcal{W}$.

To deep into the correlation among these two series, Fig. 4 shows their additive decomposition. As can be seen, both series do not have a clear increasing or decreasing trend component, with a quite stochastic behavior in case of $\mathcal{W}$. This figure also shows that the residual component of $\mathcal{W}$ is much more important than its trend and seasonal components. Whilst this residual component roughly ranges between -100,000 and 300,000, the trend one varies between 25,000 and 125,000 and the seasonal one between -40,000 and 60,000. This large difference does not occur in $\mathcal{F}_{target}$ in such a clear manner.

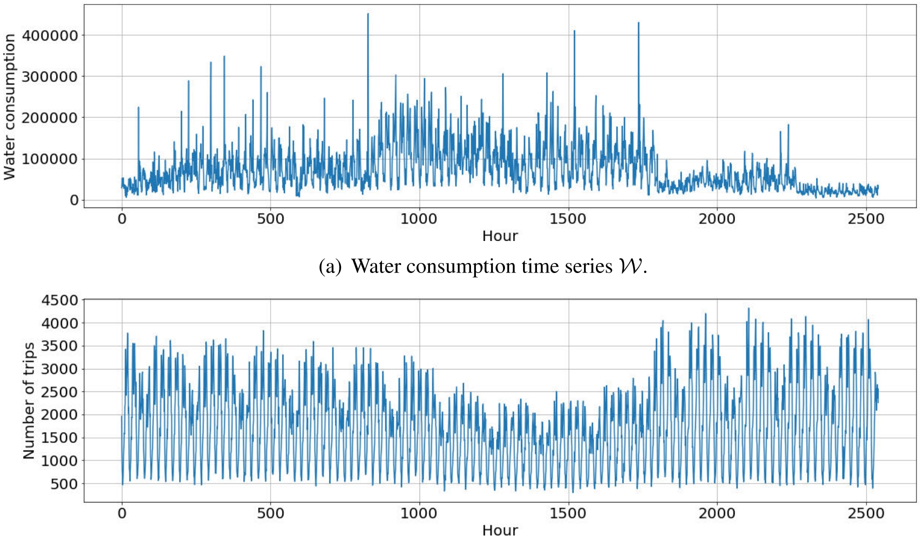
Regarding the seasonal dimension of both time series, Fig. 5 shows this dimension for a 24-hour interval. From these two dimensions it is possible to observe some interesting human patterns. In the case of the $\mathcal{W}$ component (see Fig. 5a), it can be seen three peaks at 9:00am, midday and midnight. These peaks are compatible with the time period in which people tend to stay at home having breakfast, lunch or dinner. In the case of $\mathcal{F}_{target}$, the three peaks occur at 13:00, 14:00 and 20:00 (see Fig. 5b). Again, this is compatible with the regular behavior of people going home at the end of the morning and late at the evening.

These two seasonal dimensions also explain the negative PCC observed in Table 2. Thus, whilst $\mathcal{F}_{target}$ exhibits a clear increasing trend during the afternoon and evening (see Fig. 5b), $\mathcal{W}$ mirrors such a behaviour with water consumption peaks in the morning and at night (see Fig. 5a).

Finally, the correlation scores were calculated again but only considering the seasonal component of $\mathcal{W}$ (denoted $\mathcal{W}^s$). The obtained PCC and MIS are shown in the two rightmost columns of Table 2. As can be seen, these new

TABLE 2. Correlation values between each extracted human flow and the water consumption series $\mathcal{W}$. The highest value of each column is shown in bold.

Flow	PCC($\mathcal{F}_\theta, \mathcal{W}$)	MIS($\mathcal{F}_\theta, \mathcal{W}$)	PCC($\mathcal{F}_\theta, \mathcal{W}^s$)	MIS($\mathcal{F}_\theta, \mathcal{W}^s$)
$\mathcal{F}_1$	-0.342	0.109	-0.434	0.358
$\mathcal{F}_{2/target}$	-0.372	**0.139**	-0.520	0.609
$\mathcal{F}_3$	-0.374	0.123	**-0.530**	0.680
$\mathcal{F}_4$	-0.141	0.067	-0.019	0.338
$\mathcal{F}_5$	-0.190	0.068	-0.091	0.636
$\mathcal{F}_6$	-0.182	0.098	-0.088	0.674
$\mathcal{F}_7$	-0.273	0.082	-0.254	0.514
$\mathcal{F}_8$	-0.256	0.100	-0.161	0.661
$\mathcal{F}_9$	-0.197	0.080	-0.108	0.702
$\mathcal{F}_{10}$	-0.271	0.065	-0.256	0.487
$\mathcal{F}_{11}$	-0.291	0.120	-0.267	0.668
$\mathcal{F}_{12}$	-0.260	0.097	-0.260	**0.737**
$\mathcal{F}_{13}$	-0.203	0.089	-0.499	0.497
$\mathcal{F}_{14}$	-0.284	0.130	-0.278	0.701
$\mathcal{F}_{15}$	-0.262	0.069	-0.301	0.683
$\mathcal{F}_{16}$	-0.366	0.133	-0.499	0.497
$\mathcal{F}_{17}$	-0.218	0.062	-0.113	0.634
$\mathcal{F}_{18}$	-0.187	0.082	-0.100	0.680
$\mathcal{F}_{19}$	**-0.380**	0.119	-0.509	0.550
$\mathcal{F}_{20}$	-0.294	0.114	-0.184	0.491
$\mathcal{F}_{21}$	-0.255	0.100	-0.183	0.603



(a) Water consumption time series $\mathcal{W}$.

(b) Target human flow $\mathcal{F}_{target}$.

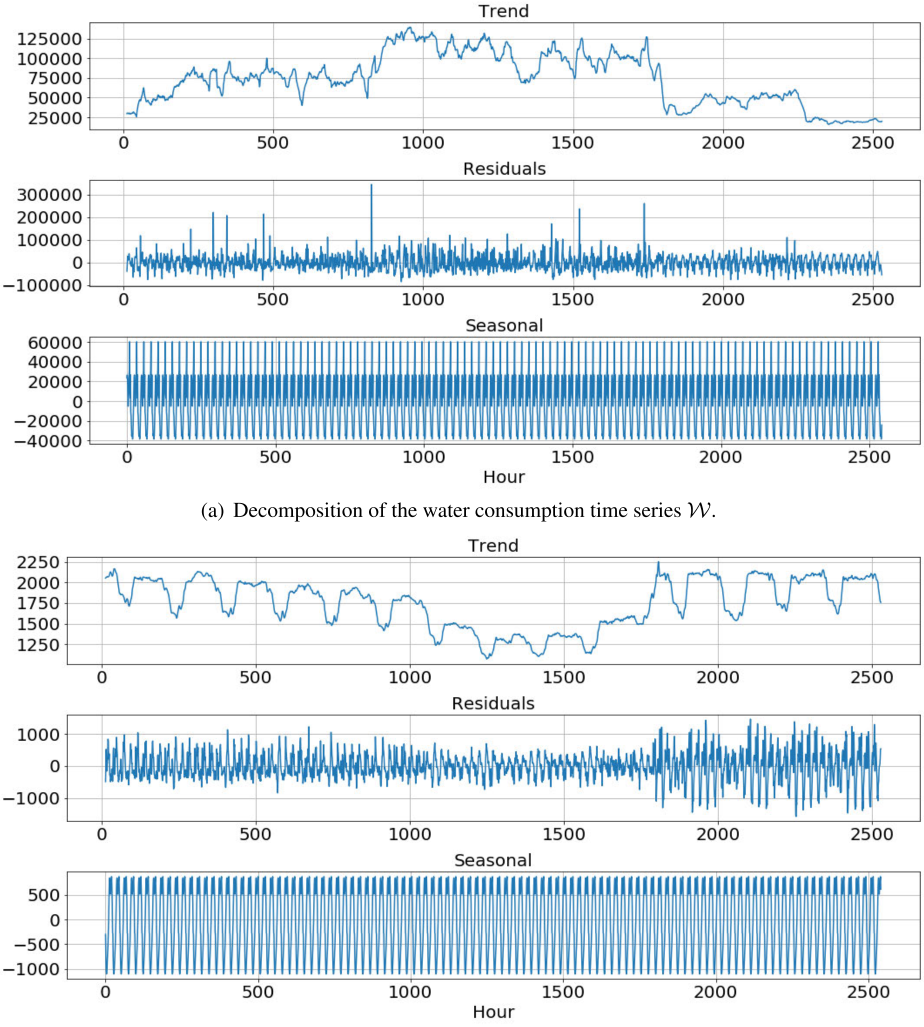
FIGURE 3. Raw time series of the target mobility flow and the water-consumption series.

values are notably higher than the ones calculated by considering the entire $\mathcal{W}$ series. For example, the PCC of $\mathcal{F}_{target}$ ($\mathcal{F}_2$ in Table 2) increases to 0.520 when compared to $\mathcal{W}^s$. The same occurs with the MIS metric increasing from 0.139 to 0.609. Therefore, the prediction mechanism to be developed must take into account this high correlation between the seasonal dimension of $\mathcal{W}$ and the target human flow $\mathcal{F}_{target}$.

D. WATER DATA CURATION

Given the noisy nature of $\mathcal{W}$ observed in the correlation study, a smooth filtering was performed by means of a Kalman Filter (KF) [37]. This method has been widely used for time series smoothing in a large range of domains [38]–[40]. KF provides a sequential, unbiased, and minimum error variance estimate that works well for discrete-time filtering problems where the underlying physical phenomenon is modeled as a discrete-time process [41]. As a result, a new smooth water-consumption time series $\mathcal{W}_{smooth}$ was generated as shown in Fig. 6a.

This new time series reduces the impact that the residual dimension had on $\mathcal{W}$ and it makes more relevant the seasonal part of the series. As observed from Fig. 6b, the residual dimension now ranges from -50,000 to 75,000 whereas in



(a) Decomposition of the water consumption time series $\mathcal{W}$.

(b) Decomposition of the human flow time series $\mathcal{F}_{target}$.

FIGURE 4. Decomposition of the target human mobility flow and the water consumption time series.

the original dataset its range of values was notably larger (−100,000, 300,000) (see Fig. 4a). Likewise, the seasonal dimension range shifts from (−40,000, 60,000) to (−20,000, 20,000). Finally, this new smoothed data is used in the following stages of the work.

It is worth mentioning that, in operational terms, the smooth step performed by the Kalman Filter can be done in real time because its recursive structure does not require to store observations or past estimates. Therefore, it is feasible to apply this filter to the raw water consumption data before it is processed by the proposed forecasting model.

IV. DESIGN OF A MOBILITY PREDICTION WITH WATER CONSUMPTION AND REGION-BASED FLOWS

This section describes in detail the human mobility predictor based on the target flow and water consumption time series described in sec. III-C and sec. III-D.

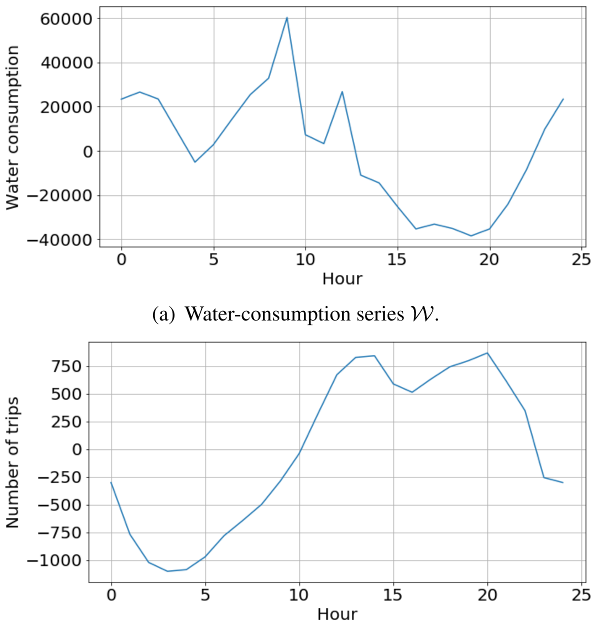
A. PROBLEM FORMULATION

The human-mobility prediction problem that the present work focuses on can be formulated as the following regression problem,

Given the hour $h \in \langle 0, .., 23\rangle$, the number of incoming home trips during the last h_{prev} previous hours in the MA m_A, $\mathcal{F}^h_{target} = \langle f^h_{target}, f^{h-1}_{target}, .., f^{h-h_{prev}}_{target}\rangle$ and the smoothed water consumption from a close residential area during the same hours $\mathcal{W}^h_{smooth} = \langle w^h_{smooth}, w^{h-1}_{smooth}, .., w^{h-h_{prev}}_{smooth}\rangle$, **Find** a mapping function $\mathcal{P}$,

$$\mathcal{P}(\mathcal{F}^h_{target}, \mathcal{W}^h_{smooth}) \rightarrow f^{h+T}_{target}$$

where f^{h+T}_{target} is the sheer number of incoming home trips in m_A in the $h + T$ hour being T the time horizon of the prediction ($T \geq 1$).



(a) Water-consumption series $\mathcal{W}$.

(b) Human-flow series $\mathcal{F}_{target}$.

FIGURE 5. Seasonal dimension of $\mathcal{W}$ and $\mathcal{F}_{target}$ for a 24-hour interval.

B. PROPOSED MODEL

Given the sequence nature of the two input sources of our approach ($\mathcal{F}_{target}$ and $\mathcal{W}_{smooth}$), we have used a Gated Recurrent Unit (GRU) model to solve the prediction problem formulated in the previous section. This is a foremost variant of Recurrent Neural Networks (RNN) [42]. In brief, GRU models are able to learn short-term and long term patterns in sequences of data. Unlike other foremost RNN models like Long Short-Term Memory (LSTM) models, a GRU model has a slightly simpler structure which makes it faster to train [43].

Fig. 7 depicts the general architecture of the GRU model and the inner structure of its cells. As we can see, a GRU model just follows the composition of a regular RNN. Concerning the structure of the cells, Fig. 7b shows that they make use of a gated mechanism to memorize long-term patterns in the target sequence. Thus, a cell receives as input the current input vector $x(t)$ and the previous state vector $h(t-1)$. Then, the cell generates the associated output $y(t)$ which is also the state vector $h(t)$ of the next cell.

More in detail, the cell comprises three different gates, the update $z(t)$, the reset $r(t)$ and memory-content $g(t)$. The computations of each gate are as follows,

$$z_{(t)} = \sigma(W_z\, x_t + U_z\, h_{(t-1)} + b_z) \tag{1}$$

$$r_{(t)} = \sigma(W_r\, x_t + U_t\, h_{(t-1)} + b_t) \tag{2}$$

$$g_{(t)} = tanh(W_g\, x_t + U_g\, (r_{(t)} \otimes h_{(t-1)}) + b_g) \tag{3}$$

$$y_{(t)} = h_{(t)} = z_{(t)} \otimes h_{(t-1)} + (1 - z_{(t)}) \otimes g_{(t)} \tag{4}$$

where $W_{\{z,r,g\}}$ are the weight matrices for the input x_t, $U_{\{z,r,g\}}$ are the weight matrices for the connections to the previous short-term state $h(t-1)$ and $b_{\{z,r,g\}}$ are the bias terms of each layer.

V. EVALUATION OF THE PREDICTOR

In order to evaluate the suitability of our proposal, two different GRU models were generated. One was fed with $\mathcal{F}_{target}$ and $\mathcal{W}$ whereas the other only took as input the human-mobility flow. This way, it was possible to assess the actual benefit of enriching a prediction model with the water consumption dataset.

A. METRICS

Regarding the metrics to perform the aforementioned evaluation, the Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) [44] are three of the most common metrics used to measure accuracy for continuous variables. They are suitable for model comparisons because they express average model prediction error in the units of the variable of interest. Their definition is as follows:

$$\begin{aligned} MSE &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i,)^2 \\ RMSE &= \sqrt{MSE}, \\ MAE &= \frac{\sum_{i=1}^{n}|y_i - \bar{y}_i|}{n}, \end{aligned}$$

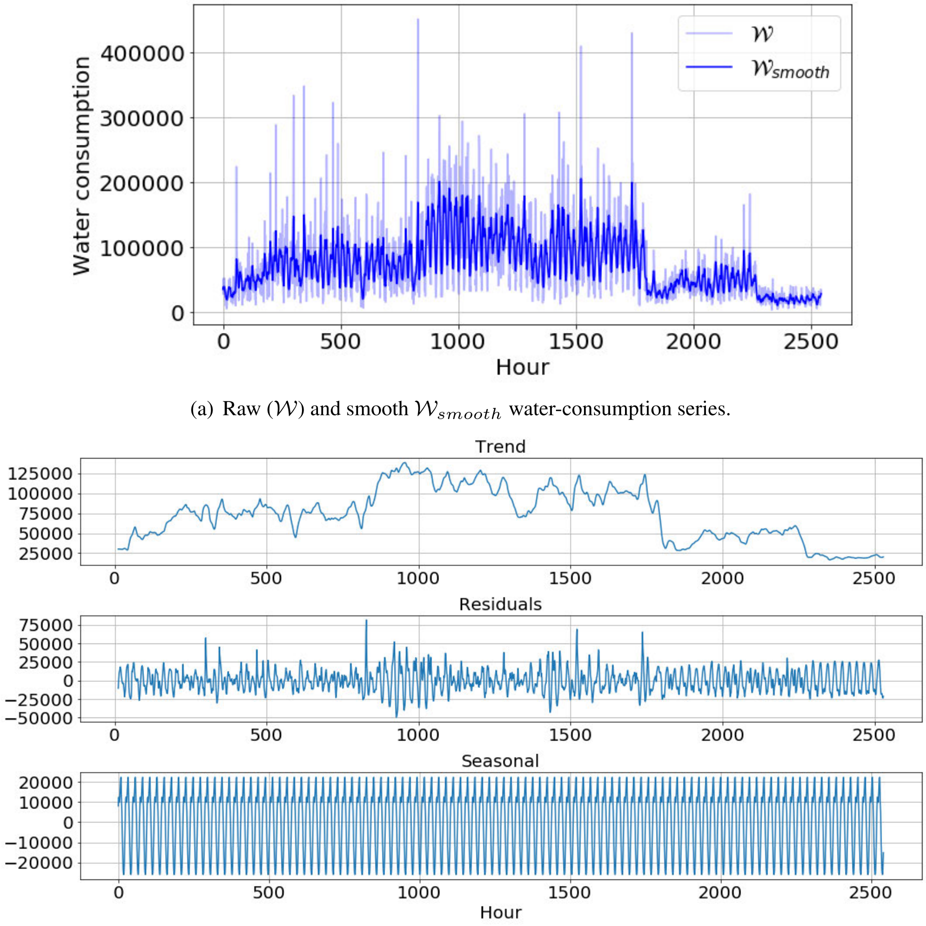
where, for our use case, y_i is the real number of incoming home trips, $\bar{y}_i$ is the predicted number of trips and n is the number of observations.

Furthermore, we complement the metrics with the coefficient of variance of the RMSE. The Coefficient of Variation of the RMSE (CVRMSE) is a non-dimensional measure calculated by dividing the RMSE of the predicted number of trips by the mean value of the actual number of trips. For example, a CVRMSE value of 5% would indicate that the mean variation in the actual number of trips which is not explained by the prediction model is 5% of the mean value of the actual number of target trips [45]. Similarly, the Mean Average Prediction Error (MAPE) metric expresses the average absolute error as a percentage. They are calculated as follows:

$$\begin{aligned} CVRMSE &= \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}}{\bar{y}} \times 100, \\ MAPE &= \frac{1}{n}\sum_{i=1}^{n}|\frac{y_i - \bar{y}_i}{y_i}| \times 100. \end{aligned}$$

B. SINGLE MODEL COMPARISON (GRU_{single})

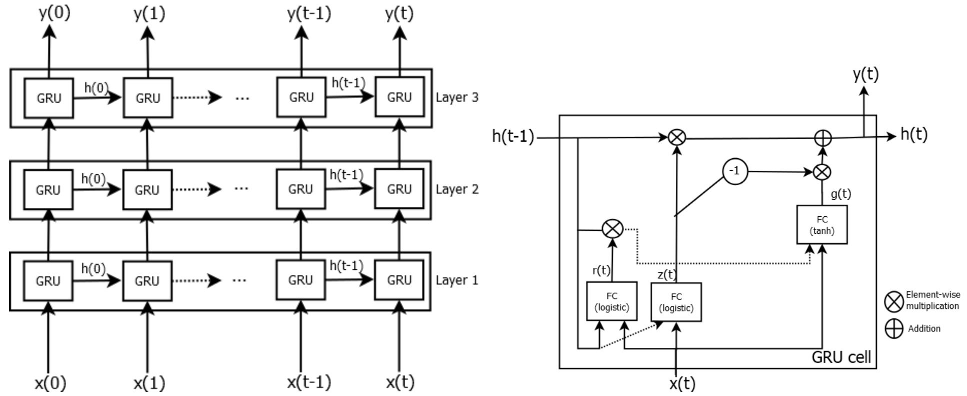
In this first evaluation, the two aforementioned GRU models were configured in the same way. This allowed to compare the actual benefit of enriching an initial model with the water consumption data without adding complexity to the model.



(a) Raw ($\mathcal{W}$) and smooth $\mathcal{W}_{smooth}$ water-consumption series.

(b) Decomposition of $\mathcal{W}_{smooth}$.

FIGURE 6. **Water-consumption data smoothing.**



(a) Example of GRU architecture with three layers unrolled thought time.

(b) Gated structure of a GRU cell. FC stands for Fully Connected.

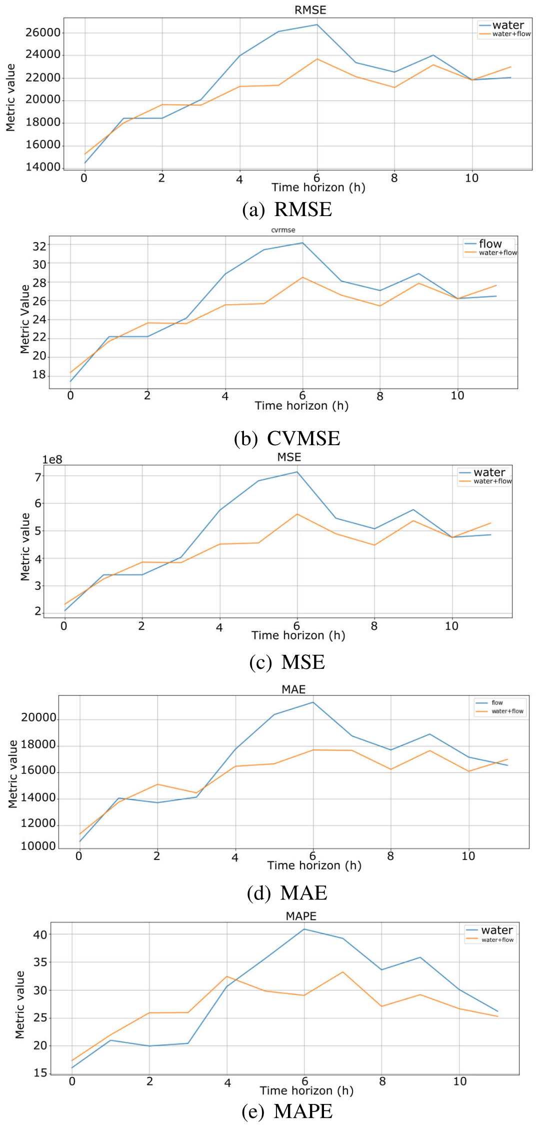
FIGURE 7. **Gated Recurrent Unit (GRU) general structure.**

In that sense, the column GRU_{single} of Table 3 shows the configuration of both models in this first comparison.

Fig. 8 shows the metrics of the two models when different time horizons T from 1 to 12h are used. From these plots, it is possible to see that adding the water consumption as input clearly improved the results of the model. This is particularly remarkable for long-term predictions when the target time horizon is larger than 3 hours ($T \geq 3h$). For example, according to Fig. 8b, the GRU model fed with the flow and water consumption sources ($GRU_{single}(\mathcal{F}_{target}, \mathcal{W}_{smooth})$) had a 25.7 CVRMSE when it came to predict the incoming number of home trips of m_A 4 hours in advance ($T = 4h$).

TABLE 3. Parameters of the models. The *Num. of epochs* indicate the epoch at which the training finished given the early-stopping policy.

Parameter	GRU_{single}	$GRU_{best}(\mathcal{F}_{target})$	$GRU_{best}(\mathcal{F}_{target}, \mathcal{W}_{smooth})$
Training rate	70%		
Loss function	Mean Squared Error (MSE)		
Activation function	Hyperbolic tangent function		
Batch size	32	32	64
Learning factor	0.001	0.01	0.01
Optimizer	Adam	root mean squared (RMS)	root mean squared (RMS)
Num. of layers	4	3	3
Num of cells per layer	32	128	128
Num. of epochs	86	32	26



(a) RMSE

(b) CVMSE

(c) MSE

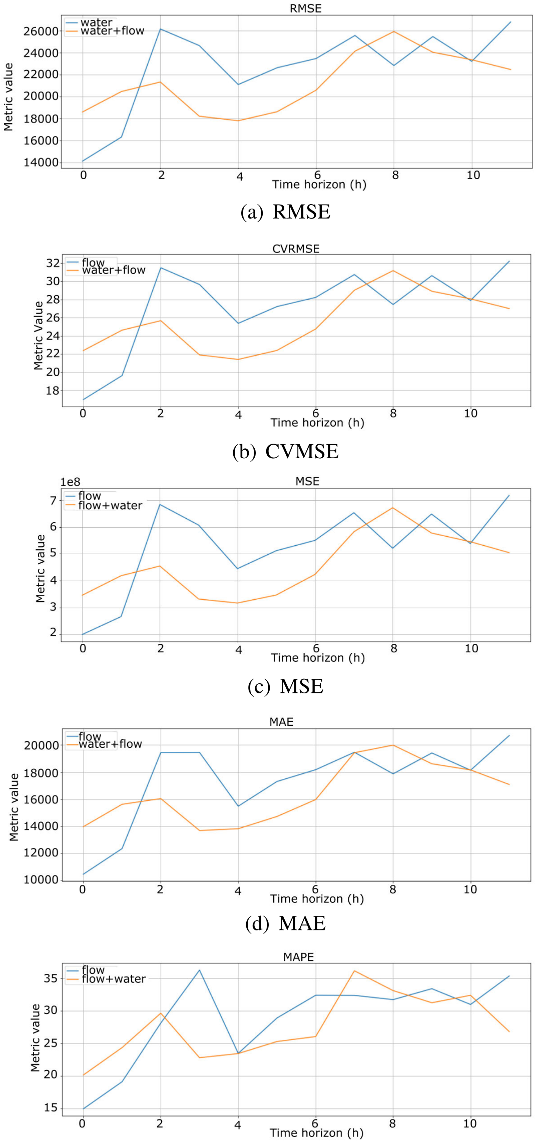
(d) MAE

(e) MAPE

FIGURE 8. Metric values for the GRU_{single} comparison for different time horizons T.

However, the same model only taking $\mathcal{F}_{target}$ as input ($GRU_{single}(\mathcal{F}_{target})$) has a CVRMSE value of 28.3. A similar pattern was observed in the other four metrics.

Finally, the two first rows of Table 4 include the mean and standard deviation of the metrics for the single models considering all the time horizons. As we can see, the model enriched with the water consumption data outperforms the one based solely on human flows in all the metrics.



(a) RMSE

(b) CVMSE

(c) MSE

(d) MAE

(e) MAPE

FIGURE 9. Metric values for the GRU_{best} comparison for different time horizons T.

C. BEST MODEL COMPARISON (GRU_{best})

The second experiment focused on fine tuning a different GRU model for each input configuration. This way, it was possible to actually assess the impact of using water consumption for human mobility prediction when the best model is used in each case. In that sense, the two rightmost columns in Table 3 show the actual configuration of both models.

According to Fig. 9, this second experiment confirmed the benefit using the water data with respect the model only fed with the target human flow. For example, Fig. 9b shows that the CVRMSE of the model based on the two inputs ($GRU_{best}(\mathcal{F}_{target}, \mathcal{W}_{smooth})$) was 21.9 when it came to predict the incoming flow of home trips 3 hours ahead ($T = 3h$) whereas the model based on human flows, $GRU_{best}(\mathcal{F}_{target})$, achieved 29.8 as value. Furthermore, the two last rows of

TABLE 4. Mean and standard deviation of the metrics for the different experiments. The best metric value is marked in bold.

Model	Metric				
	MAE	MSE	RMSE	CVRMSE	MAPE
$GRU_{single}(\mathcal{F}_{target})$	17,346.661 (3112.264)	5.285×10^8 (±1.601×10^8)	22,685.122 (±3,893.455)	27.285 (±4.676)	28.910 (±6.565)
$GRU_{single}(\mathcal{F}_{target}, \mathcal{W}_{smooth})$	16,415.441 (±2,222.796)	4.597×10^8 (±1.165×10^8)	21,284.753 (±2,700.971)	25.601 (±3.239)	27.606 (±4.854)
$GRU_{best}(\mathcal{F}_{target})$	16,757.739 (±3,061.191)	4.873×10^8 (±1.465×10^8)	21,820.523 (±3,499.267)	26.245 (±4.203)	29.105 (±8.272)
$GRU_{best}(\mathcal{F}_{target}, \mathcal{W}_{smooth})$	**15,838.033 (±1,894.534)**	**4.388**×10^8 **(±9.432**×10^8**)**	**20,822.338 (±2,396.228)**	**25.045 (±2.873)**	**26.961 (±4.343)**

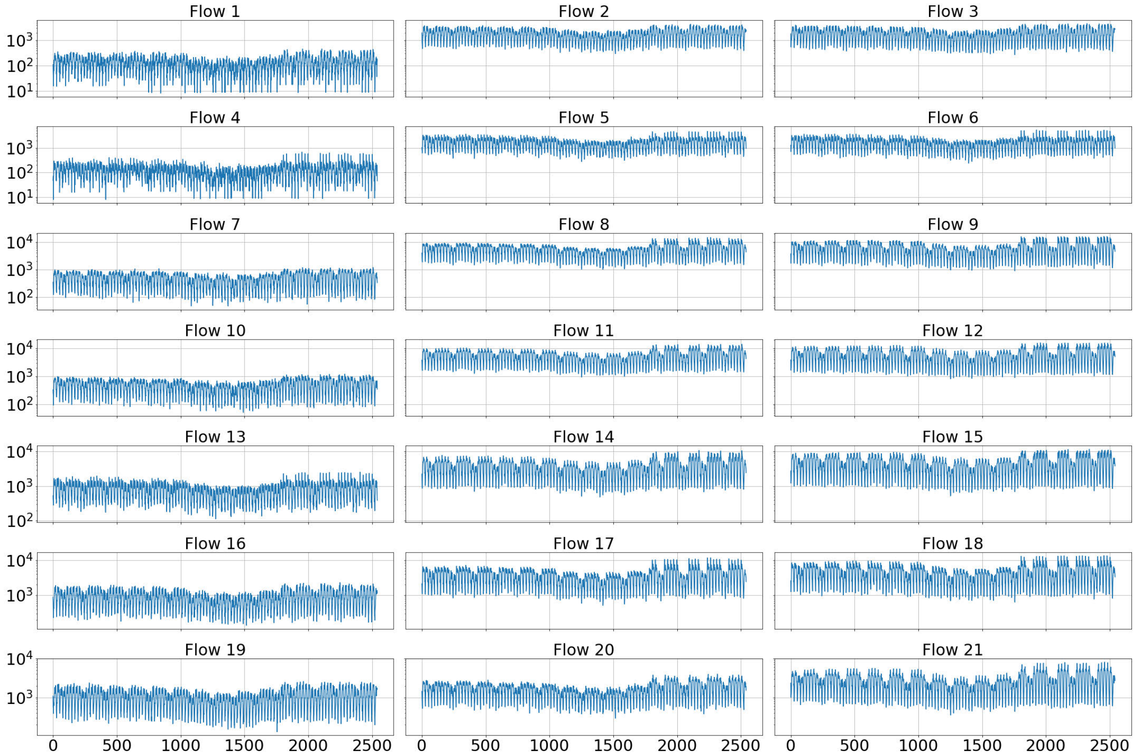


FIGURE 10. Time series of all the generated human flows.

Table 4 also indicated that, on average, the best model with water and human flow data outperformed the one only relying on mobility data regardless of the time horizon.

To conclude, it is important to remark that in the two experiments the most important improvement in the prediction accuracy was for time horizons between 3 and 7 hours (see Figs. 8 and 9). This makes sense as the domestic water consumption might capture anomalies in the human presence at home in the target MA that are not captured by the human-mobility flow. However, these anomalies would impact the incoming flow of home trips some hours later (after people leave the MA and then go back). Actually, bearing in mind the seasonal dimensions of both time series in Fig. 5, the differences between the three peaks of both series fall within the aforementioned range of time-horizon values, 09:00 vs 13:00, 12:00 vs 14:00 and 00:00 vs 20:00.

VI. CONCLUSION

The analysis of people's mobility patterns is a key factor in modern societies. Actually, the mining of such patterns is an instrumental point to develop intelligent solutions in a wide range of domains. Existing technologies, such as WiFI or Bluetooth, allow geolocation of citizens, but are constrained by regulatory restrictions. In this paper, we propose a non-invasive methodology to analyse, in an anonymous and aggregated way, the coarse-grained mobility patterns of individuals.

Our results demonstrate that the use of low-resolution mobility data together with the use of household water-consumption data can facilitate the task of prediction of mobility at a given location over horizons between 3-7 hours. As we can see, the use of essential supplies such as water or electricity, on an aggregated basis, provides anonymous patterns that are geolocated by coarse-grained regions of interest, and can measure mobility within those regions. This opens up a range of opportunities to predict mobility patterns in particular locations without the need to invade the privacy of individuals.

Finally, future work will focus on aggregating other contextual sources to the model. For example, the information of holidays and weather conditions related to the target areas might improve the prediction accuracy of the models.

APPENDIX A
GENERATED HUMAN FLOWS

Fig. 10 shows all the trip flows generated for the present study.

REFERENCES

[1] G. Solmaz and D. Turgut, "A survey of human mobility models," *IEEE Access*, vol. 7, pp. 125711–125731, 2019.

[2] J. Guo, S. Zhang, J. Zhu, and R. Ni, "Measuring the gap between the maximum predictability and prediction accuracy of human mobility," *IEEE Access*, vol. 8, pp. 131859–131869, 2020.

[3] W. Xi, T. Pei, Q. Liu, C. Song, Y. Liu, X. Chen, J. Ma, and Z. Zhang, "Quantifying the time-lag effects of human mobility on the COVID-19 transmission: A multi-city study in China," *IEEE Access*, vol. 8, pp. 216752–216761, 2020.

[4] J. Wang, X. Kong, A. Rahim, F. Xia, A. Tolba, and Z. Al-Makhadmeh, "IS$_2$Fun: Identification of subway station functions using massive urban data," *IEEE Access*, vol. 5, pp. 27103–27113, 2017.

[5] P. Castrogiovanni, E. Fadda, G. Perboli, and A. Rizzo, "Smartphone data classification technique for detecting the usage of public or private transportation modes," *IEEE Access*, vol. 8, pp. 58377–58391, 2020.

[6] Q. Liu, X. Zheng, H. E. Stanley, F. Xiao, and W. Liu, "A spatio-temporal co-clustering framework for discovering mobility patterns: A study of manhattan taxi data," *IEEE Access*, vol. 9, pp. 34338–34351, 2021.

[7] M. Batran, M. Mejia, H. Kanasugi, Y. Sekimoto, and R. Shibasaki, "Inferencing human spatiotemporal mobility in greater Maputo via mobile phone big data mining," *ISPRS Int. J. Geo-Inf.*, vol. 7, no. 7, p. 259, Jun. 2018.

[8] H. Ullah, W. Wan, S. A. Haidery, N. U. Khan, Z. Ebrahimpour, and ., "Spatiotemporal patterns of visitors in urban green parks by mining social media big data based upon WHO reports," *IEEE Access*, vol. 8, pp. 39197–39211, 2020.

[9] M. von Mörner, "Application of call detail Records–Chances and obstacles," *Transp. Res. Procedia*, vol. 25, pp. 2233–2241, Jan. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2352146517307366

[10] H. F. Chan, A. Skali, and B. Torgler, "A global dataset of human mobility," Center Res. Econ., Manage. Arts (CREMA), Washington, DC, USA, Work. Paper 2020-04, 2020.

[11] G. Barlacchi, M. De Nadai, R. Larcher, A. Casella, C. Chitic, G. Torrisi, F. Antonelli, A. Vespignani, A. Pentland, and B. Lepri, "A multi-source dataset of urban life in the city of Milan and the province of trentino," *Sci. Data*, vol. 2, no. 1, pp. 1–15, Dec. 2015.

[12] S. Chang, E. Pierson, P. W. Koh, J. Gerardin, B. Redbird, D. Grusky, and J. Leskovec, "Mobility network models of COVID-19 explain inequities and inform reopening," *Nature*, vol. 589, p. 87, Jan. 2020.

[13] B. Fu, N. Damer, F. Kirchbuchner, and A. Kuijper, "Sensing technology for human activity recognition: A comprehensive survey," *IEEE Access*, vol. 8, pp. 83791–83820, 2020.

[14] M. Rizwan, W. Wan, and L. Gwiazdzinski, "Visualization, spatiotemporal patterns, and directional analysis of urban activities using geolocation data extracted from LBSN," *ISPRS Int. J. Geo-Inf.*, vol. 9, no. 2, p. 137, Feb. 2020.

[15] W. Mungthanya, S. Phithakkitnukoon, M. G. Demissie, L. Kattan, M. Veloso, C. Bento, and C. Ratti, "Constructing time-dependent origin-destination matrices with adaptive zoning scheme and measuring their similarities with taxi trajectory data," *IEEE Access*, vol. 7, pp. 77723–77737, 2019.

[16] F. Terroso-Saenz, A. Muñoz, and F. Arcas, "Land-use dynamic discovery based on heterogeneous mobility sources," *Int. J. Intell. Syst.*, vol. 36, no. 1, pp. 478–525, Jan. 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/int.22307

[17] J. Bao, X. Shi, and H. Zhang, "Spatial analysis of bikeshare ridership with smart card and POI data using geographically weighted regression method," *IEEE Access*, vol. 6, pp. 76049–76059, 2018.

[18] K. Smolak, B. Kasieczka, W. Fialkiewicz, W. Rohm, K. Siła-Nowicka, and K. Kopańczyk, "Applying human mobility and water consumption data for short-term water demand forecasting using classical and machine learning models," *Urban Water J.*, vol. 17, no. 1, pp. 32–42, Jan. 2020, doi: 10.1080/1573062X.2020.1734947.

[19] K. Smolak, B. Kasieczka, K. Sila-Nowicka, K. Kopanczyk, W. Rohm, and W. Fialkiewicz, "Urban hourly water demand prediction using human mobility data," in *Proc. IEEE/ACM 5th Int. Conf. Big Data Comput. Appl. Technol. (BDCAT)*, Dec. 2018, pp. 213–214.

[20] S. Shahriari, M. Ghasri, S. A. Sisson, and T. Rashidi, "Ensemble of ARIMA: Combining parametric and bootstrapping technique for traffic flow prediction," *Transportmetrica A, Transp. Sci.*, vol. 16, no. 3, pp. 1552–1573, Jan. 2020.

[21] B. Alsolami, R. Mehmood, and A. Albeshri, "Hybrid statistical and machine learning methods for road traffic prediction: A review and tutorial," in *Smart Infrastructure and Applications* (EAI/Springer Innovations in Communication and Computing), R. Mehmood, S. See, I. Katib, and I. Chlamtac, Eds. Cham, Switzerland: Springer, 2020, pp. 115–133.

[22] S. Kwak and N. Geroliminis, "Travel time prediction for congested freeways with a dynamic linear model," *IEEE Trans. Intell. Transp. Syst.*, early access, Jul. 22, 2020, doi: 10.1109/TITS.2020.3006910.

[23] V. Kulkarni, A. Mahalunkar, B. Garbinato, and J. D. Kelleher, "On the inability of Markov models to capture criticality in human mobility," in *Artificial Neural Networks and Machine Learning—ICANN 2019: Image Processing* (Lecture Notes in Computer Science), vol. 11729, I. Tetko, V. Kurkova, P. Karpov, and F. Theis, Eds. Munich, Germany: Springer, 2019, pp. 484–497.

[24] A. Kurkcu, K. Ozbay, and E. Morgul, "Evaluating the usability of geo-located Twitter as a tool for human activity and mobility patterns: A case study for NYC," in *Proc. Transp. Res. Board's 95th Annu. Meeting*, 2016, pp. 1–20.

[25] N. Pourebrahim, S. Sultana, A. Niakanlahiji, and J.-C. Thill, "Trip distribution modeling with Twitter data," *Comput., Environ. Urban Syst.*, vol. 77, Sep. 2019, Art. no. 101354.

[26] E. Alomari, I. Katib, and R. Mehmood, "Iktishaf: A big data road-traffic event detection tool using Twitter and spark machine learning," *Mobile Netw. Appl.*, pp. 1–16, Aug. 2020.

[27] S. Miyazawa, X. Song, R. Jiang, Z. Fan, R. Shibasaki, and T. Sato, "City-scale human mobility prediction model by integrating GNSS trajectories and SNS data using long short-term memory," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vols. V-4-2020, pp. 87–94, Aug. 2020.

[28] D. Kong and F. Wu, "HST-LSTM: A hierarchical spatial-temporal long-short term memory network for location prediction," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2341–2347.

[29] J. Zhao, J. Xu, R. Zhou, P. Zhao, C. Liu, and F. Zhu, "On prediction of user destination by sub-trajectory understanding: A deep learning based approach," in *Proc. 27th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2018, pp. 1413–1422.

[30] Z. Fan, X. Song, T. Xia, R. Jiang, R. Shibasaki, and R. Sakuramachi, "Online deep ensemble learning for predicting citywide human mobility," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–21, Sep. 2018.

[31] Z. Cui, K. Henrickson, R. Ke, and Y. Wang, "Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 11, pp. 4883–4894, Nov. 2020.

[32] L. Zhao, Y. Song, C. Zhang, Y. Liu, P. Wang, T. Lin, M. Deng, and H. Li, "T-GCN: A temporal graph convolutional network for traffic prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3848–3858, Sep. 2020.

[33] A. Di Mauro, A. Cominola, A. Castelletti, and A. Di Nardo, "Urban water consumption at multiple spatial and temporal scales. A review of existing datasets," *Water*, vol. 13, no. 1, p. 36, Dec. 2020.

[34] H. Barbosa, M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini, "Human mobility: Models and applications," *Phys. Rep.*, vol. 734, pp. 1–74, Mar. 2018.

[35] M. Y. A. U. S. de Estado de Transportes, "Análisis de la movilidad en España con tecnología Big Data durante el Estado de Alarma para la gestión de la crisis del COVID-19," Ministerio de Transportes, Movilidad y Agenda Urbana, Madrid, Spain, Tech. Rep., Apr. 2020.

[36] B. C. Ross, "Mutual information between discrete and continuous data sets," *PLoS ONE*, vol. 9, no. 2, Feb. 2014, Art. no. e87357.

[37] Y. Kim and H. Bang, "Introduction to Kalman filter and its applications," *Introduction Implement. Kalman Filter*, vol. 1, pp. 1–16, 2018.

[38] L. Ralaivola and F. d'Alche-Buc, "Time series filtering, smoothing and learning using the kernel Kalman filter," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jul. 2005, pp. 1449–1454.

[39] F. Avanzi, Z. Zheng, A. Coogan, R. Rice, R. Akella, and M. H. Conklin, "Gap-filling snow-depth time-series with Kalman filtering-smoothing and expectation maximization: Proof of concept using spatially dense wireless-sensor-network data," *Cold Regions Sci. Technol.*, vol. 175, Jul. 2020, Art. no. 103066.

[40] S. Tavakoli, H. Fasih, J. Sadeghi, and H. Torabi, "Kalman filter-smoothed random walk based centralized controller for multi-input multi-output processes," *Int. J. Ind. Electron., Control Optim.*, vol. 2, no. 2, pp. 155–166, 2019.

[41] S. Sarkka, A. Vehtari, and J. Lampinen, "Time series prediction by Kalman smoother with cross-validated noise density," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Jul. 2004, pp. 1653–1657.

[42] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN Encoder–Decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1724–1734.

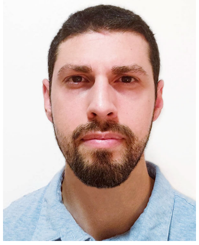
[43] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *Proc. NIPS Workshop Deep Learn.*, Dec. 2014, pp. 1–9.

[44] C. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, 2005.

[45] T. A. Reddy, N. F. Saman, D. E. Claridge, J. S. Haberl, W. D. Turner, and A. T. Chalifoux, "Baselining methodology for facility-level monthly energy use—Part 1: Theoretical aspects," in *ASHRAE Transactions*. Atlanta, GA, USA: ASHRAE, 1997, pp. 336–347.

FERNANDO TERROSO-SÁENZ received the B.S. and Ph.D. degrees in computer science from the University of Murcia, in 2009 and 2013, respectively. Since 2017, he has been an Associate Professor with the Catholic University of Murcia (UCAM). He has published more than 30 articles in international journals and conference. His research interests include smart mobility, human-generated data analysis, and mobile sensing.

ANDRÉS MUÑOZ received the B.S. degree in computer science and the Ph.D. degree in computer science from the University of Murcia, Spain, in 2005 and 2011, respectively. He is currently an Associate Professor with the Polytechnic School, Catholic University of Murcia. His research interests include Semantic Web technologies, ambient intelligence, and intelligent environments.



JULIO FERNÁNDEZ-PEDAUYE received the B.Sc. degree in computer science from the Catholic University of Murcia, in 2020. He is currently a Computer Scientist with the Universitat Politècnica de València, Spain. His current research interests include natural language processing and social computing.

JOSÉ M. CECILIA received the degrees in computer engineering from the University of Murcia, and Cranfield University, in 2005 and 2007, respectively, and the Ph.D. degree in computer architecture from the University of Murcia, in 2011. He is currently a Ramón y Cajal Research Fellow and an Associate Professor (tenure track) with the Department of Computer Engineering, Universitat Politècnica de València, Spain. His research interests include high performance computing, social sensing applications, and IA.