MASTER

Multiple ASpect TrajEctoRy management and analysis

Project Acronym	***MASTER***
Project Full Name	***M**ultiple **AS**pects **T**raj**E**cto**R**y management and analysis*
Project Number	***777695***
Deliverable Title	***Requirements for application scenarios***
Deliverable No.	***D5.1***
Contract Delivery Date	***28/02/2019 (M12)***
Actual Delivery Date	***28/02/2019(M12)***
Responsible Authors	*Nikos Pelekis (Unit Manager of UPRC)*
	This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie-Sklodowska Curie grant No 777695

DOCUMENT INFORMATION

GRANT AGREEMENT N.	777695
PROJECT ACRONYM	MASTER
PROJECT FULL NAME	Multiple Aspects Trajectory Management and Analysis
STARTING DATE (DUR.)	01/03/2018 (48 months)
ENDING DATE	28/2/2022
PROJECT WEBSITE	http://www.master-project-h2020.eu
COORDINATOR	Chiara Renso (CNR)
WORKPACKAGE N. | TITLE | START – END MONTH	WP5 | Application Scenarios | M1 - M48
WORKPACKAGE LEADER	UPRC
DELIVERABLE N. | TITLE	D5.1 | Requirements for application scenarios
RESPONSIBLE AUTHOR	Nikos Pelekis (Unit Manager of UPRC)
DATE OF DELIVERY (CONTRACTUAL)	28 February 2019 (M12)
DATE OF DELIVERY (SUBMITTED)	28 February 2019 (M12)
VERSION | STATUS	2.0
NATURE	REPORT
DISSEMINATION LEVEL	PUBLIC
AUTHORS (PARTNER)	UPRC, CNR, UVSQ, UNIVE, HUA, Thira
CONTRIBUTORS	Nikos Pelekis (UPRC), Alessandra Raffaetà (UNIVE), Chiara Renso (CNR),

	Konstantinos Tserpes (HUA), Karine Zeitouni (UVSQ), Cristina Muntean (CNR), Ida Mele (CNR), Vinicius Monteiro de Lira (CNR), Raffaele Perego (CNR), Fabio Pranovi (UNIVE), Elisabetta Russo (UNIVE), Marta Simeoni (UNIVE), Lorenzo Gabrielli (CNR), Christallia Papaoikonomou (Thira)

ACRONYM LIST

MASTER	Multiple Aspects Trajectory Management and Analysis
ICT	Information and Communication Technologies
ISTI	Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo"
CNR	Consiglio Nazionale delle Ricerche
UNIVE	Ca' Foscari University of Venice
UVSQ	University of Versailles Saint-Quentin
UPRC	University of Pireaus Research Center
HUA	Harokopio University of Athens
PUC	Pontificial University of Rio de Janeiro
DAL	Dalhousie University
THIRA	Municipality of Thira
ER	Experienced Researcher
ESR	Early Stage Researcher
ACTV	Azienda del Consorzio Trasporti Veneziano (Company responsible for public transportation in province of Venice)
AIS	Automatic Identification System (AIS)

TABLE OF CONTENTS

1. INTRODUCTION

The objective of this deliverable, that is linked to WP5, is "to include the results of the requirement analysis for the scenarios and detail application questions and data available" as stated in the MASTER Grant Agreement Annex 1 Part A pages 18-19. This activity has been carried out during secondments linked to WP5 and parallel activities of project partners. We have executed 9,97 person months during which secondees have collected requirements and information about the datasets available at the partners relative to the three application scenarios: tourism, sea monitoring and transportation. The results are reported in this document.

Each chapter is structured into three main sections. The first section introduces the scenario, the second section reports the application description with general requirements, the third section describes the datasets available or to be collected.

The work done in preparing this deliverable is mainly connected to the secondments linked to Working Package 5 "Application Scenarios" for the requirement analysis and datasets.

WP5 has the following four objectives:

1. **to understand the application needs and perform the requirements analysis driving the design of the methods developed in WP3 and WP4**. This objective has been mainly reached, during the first year, by secondments to Thira, PUC, DAL. This objective has been fully achieved relatively to the first year;

2. **host the interaction between academic and non-academic partners to create awareness and best practices on the non-academic world needs, thus increasing the potential in career developing especially in ESRs**. This objective is reached, during the first year, by the secondments to Thira. This objective has been fully achieved relatively to the first year;

3. **test the developed techniques in real-world scenarios possibly on data owned by the non-academic partner and on the basis of their actual daily problems**. This objective has been reached, during the first year, from the secondments to DAL. This objective has been fully achieved relatively to the first year;

4. **to develop software prototypes to facilitate the interaction and transfer of knowledge among academic and non-academic partners.** This objective is to be reached by developing the research prototypes to be reported in D5.3 and D5.4 due at M36 and M48 This objective has not be achieved since it was not planned for the first year but only at M36.

The current deliverable D5.1 fulfills the first, second and third objectives w.r.t. to the progress to the goals of **MS1**. The deliverable D5.1 will feed deliverable D5.2 (M24-**MS3**) that will revise the application scenarios w.r.t. the goals of MS3. Indeed, deliverable D5.2 will identify those requirements and datasets that better fit for the research prototypes that will be objective of D5.3 (M36-**MS6**) and D5.4 (M48-**MS7**). These two deliverables will provide the preliminary and the final reports on application scenarios and software prototypes, respectively.

At the time this deliverable is submitted, we have no published research papers linked to WP5.

For the sake of completeness, and benefit of the consortium partners, we also added details on the datasets provided by partners which are not explicitly linked to a secondment.

In the Grant Agreement of MASTER, Annex 1 Part B we introduced the main datasets, in part general purpose and in part application specific, as summarized below.

General Human Movements

The GEOLIFE trajectories is a public dataset, therefore not provided by partners, and it is free for download from the following link [https://www.microsoft.com/en-us/research/project/geolife-building-social-networks-using-human-location-history/].

The Social Media data

The Flickr and Twitter datasets have been collected for the Tourism case study and they are reported in Chapter 2 page 17.

The Linked Open Data

The Linked Open Data (LOD) datasets are open data available freely online that can be downloaded based on the application needs as for example the Points of Interest (PoIs) in a given geographical area. Examples of these kinds of data are reported in tourism use case at page 16 as OpenStreetMap, and in transportation scenario at page 32 as Points of Interest in Rio de Janeiro.

Tourism Data

Santorini datasets are described in Chapter 2 starting from page 10.

Sea Monitoring Data

These datasets are described in Chapter 3 starting from page 24. Specifically, AIS vessel data provided by Dalhousie partner is described at page 24. Adriatic Sea data provided by Ca' Foscari University of Venice is described at page 27.

Public Transportation Data

The Public Transport Data is described in Chapter 4. The Rio bus data is described at page 31 and the data from ACTV, the main Public Transport Company in Venice, is illustrated at page 33.

At M12, the number of secondments linked to WP5 is 9, listed below in table 1. According to the Researcher Declarations submitted to Sygma system, the total number of person months is 9,97.

Researchers from CNR, UVSQ and UNIVE have been seconded to Thira, essentially working on task 5.1 on tourism, whose results are reported in Chapter 2 of the present deliverable. Researchers from HUA and UNIVE have been seconded to DAL partner to work on Task 5.2 on sea monitoring presented in Chapter 3. Researchers from UVSQ and CNR have been seconded to PUC to work on Task 5.3 transportation scenario reported in Chapter 4.

The objectives of WP5 that have been fully achieved by this deliverable, relatively to the first year of the project, are the number 1, 2 and 3.

The role of CNR researchers has been to work on Task 5.1 tourism and T5.3 transportation collecting requirements and datasets.

The role of Thira staff has been to provide application requirements and datasets to secondees for T5.1 tourism scenario.

The role of UVSQ researchers has been to collect requirements and datasets for Task 5.1 tourism and T5.3 transportation.

The role of UNIVE researchers has been to work on Task 5.1 Tourism, collecting requirements and datasets from Thira partner, Task 5.2 sea monitoring collecting requirements and datasets from DAL partner and as well as providing sea monitoring datasets and T5.3 transportation providing requirements and datasets.

The role of HUA researchers has been to work on Task 5.2 sea monitoring collecting requirement and datasets.

The role of PUC researchers has been to provide T5.3 transportation collecting requirements and datasets.

The role of DAL researchers has been to provide T5.2 sea monitoring collecting requirements and datasets.

The outcome of this deliverable is to feed the Deliverable D5.2 "Revision of the application scenarios" due at M24 that is a revision of the current one about application requirements and datasets and that will identify the datasets and the requirements that will be considered for the development of the research prototypes due at M36 and M48 as deliverables D5.3 and D5.4.

Table 1: Secondments linked to WP5 Application scenarios from the Sygma system

RD N.	Task	Secondee Name	Fellow ID	Sending Institution	Hosting Institution	From	To	PM
3	T5.1	Raffaele Perego	3	Consiglio Nazionale Delle Ricerche	Dimos Thiras	4/30/2018	5/31/2018	1,07
4	T5.1	Chiara Renso	4	Consiglio Nazionale Delle Ricerche	Dimos Thiras	4/30/2018	5/31/2018	1,07
5	T5.1	Karine Zeitouni	5	Universite De Versailles Saint-quentin-en-yvelines.	Dimos Thiras	5/6/2018	5/31/2018	0,87
6	T5.1	Andrea Marin	10	Universita Ca' Foscari Venezia	Dimos Thiras	6/25/2018	7/12/2018	0.60
7	T5.2	Iraklis Varlamis	9	Harokopio University	Dalhousie University	7/1/2018	8/31/2018	2
8	T5.2	Konstantinos Tserpes	8	Harokopio University	Dalhousie University	7/1/2018	8/31/2018	2
11	T5.2	Elisabetta Russo	13	Universita Ca' Foscari Venezia	Dalhousie U niversity	7/27/2018	8/29/2018	1,10
12	T5.3	Karine Zeitouni	5	Universite De Versailles Saint-quentin-en-yvelines	Faculdades Catolicas Associacao Sem Fins Lucrativos	8/18/2018	9/2/2018	0,53
18	T5.3	Vinicius Cesar Monteiro de Lira	7	Consiglio Nazionale Delle Ricerche	Faculdades Catolicas Associacao Sem Fins Lucrativos	12/8/2018	12/29/2018	0,73

2. TOURISM SCENARIO

This chapter of the deliverable is related to the activity of task 5.1. about Tourism application scenario, whose responsible is CNR. The material has been mainly prepared during the secondments to Thira, but also integrated during parallel activities of project partners.

The secondments that contributed to this chapter are linked to the Task T5.1 (secondments to Thira partner) and they are: Chiara Renso (CNR), Raffaele Perego (CNR), Karine Zeitouni (UVSQ) and Andrea Marin (UNIVE) as reported in Table 1 for a total of 3,61 PMs.

INTRODUCTION

In the MASTER Grant Agreement Annex 1 Part B we have identified a tourism use case regarding the design of techniques for the real-time monitoring of tourist flows and the recommendation of personalized itineraries in the Santorini island, Greece.

Santorini is located in the southern Aegean Sea, southeast of Greece's mainland. The island is ranked the world's top island for tourism by many magazines and travel sites and it receives every year almost 2 million visitors. The Municipality of Thira (Santorini) is one of the partners involved in the MASTER project. The Municipality is in the process of creating a Destination Marketing and Management Organization (DMMO), having the purpose of promoting the island as a worldwide tourism destination and managing effectively the huge number of visitors hosted. The DMMO will promote the development and marketing of the destination, by also focusing on convention sales, tourism marketing and services. Thira participates in MASTER aiming at exploiting the results of the project within the starting DMMO. The results in holistic trajectories analysis will support the Municipality's effort to have an accurate picture of what is really happening on the island and give personalized recommendations to the visitors.

In the present document we discuss the specific requirements of Thira collected during the secondments of CNR and UVSQ at M3 and UNIVE at M5. During the secondments, we had several discussions with the Thira Unit Manager, Christallia Papaoikonomou and two specific meetings with the participation of UPRC Unit Manager Nikos Pelekis to better understand the kinds of data that Thira can provide as well as the other relevant data that can be collected from external sources. The discussions were indeed very useful to better understand the requirements of the Municipality and how the project research activities could help the Municipality to better manage the huge tourism flows in the island. We also reached some understanding of the administrative and technological issues that have to be taken into account in order to build an automated solution for tourism monitoring. This solution should be suitable for a complex reality such as the Santorini island whose decision processes rely on a limited amount of digital information.

GENERAL REQUIREMENTS

Many tourists spend in Santorini only few days or even a single day (e.g. tourists arriving with cruises). They visit only the main Santorini attraction (the Caldera) and leave soon. The municipality wishes to encourage

tourists to stay longer in the island and to be more uniformly distributed in space and time: 1) visiting all the Points of Interest in the Island and enjoy the many hotels, shops and restaurants; 2) arriving to the island not only during the summer but also during the low season when the weather is however good in Santorini due to its lucky position in the South Aegean Sea. Thira Municipality would like to build a Tourism Observatory but encounters problems in collecting fresh data for feeding useful and actionable analyses. The goal of the Observatory includes a detailed monitoring of tourism flows to *measure* the effect of actions taken to improve them.

Previous data collection initiatives. In the project repository in a private area due to confidentiality, we have collected from Thira Municipality a document (in Greek) about the Tourism Observatory dated 2017 and authored by Prof. Yiannis Spilanis, Associate Professor from Aegean University. He prepared a report for the municipality of Thira discussing some indicators on the tourism in Santorini. From the report, it emerges that gathering data is extremely difficult for the Municipality due to political, administrative and technological reasons. Therefore, so far it was not possible to build a *real* observatory. A survey has been prepared few years ago after interviewing a sample of tourists.

The following general observations and requirements emerged from the discussions with Thira Unit Manager:

1) Preserve the **authentic style** of the Island. The Municipality does not want the island to become too technological: no Santorini Card, no ICT instrumentation that would not be possible to manage and maintain, no invasive technology.
2) **Spread the tourists in space**. Thira wishes to promote the visit of alternative locations on the island to better balance the tourists flows. The goal is not to put a limit in the arrivals but better balance the visits to the island locations, not only Caldera and Oia.
    a) Thira has a schedule of the daily arrival of cruise ships (from a system for berth allocation of cruise ships) and this allocation allows to keep the maximum number of daily cruise visitors to 8.000 people.
    b) Thira tried to find agreements with local travel companies to schedule the itineraries of buses of cruise tourists to avoid congestions, but so far it does not work very well because all cruises sell the same offers (e.g., Oia at the sunset).
    c) Diverge tourists flows to different points for sunset, not all the tourists at Oia. There are other wonderful places for enjoying the sunset.
3) **Spread the tourists in time**. Thira wishes to promote the visit of Santorini during the low season. *Santorini All Year Around* is an initiative for encouraging tourists to visit the island during winter and local tourism services to remain open in the low season. The result of the initiative is fairly good, but there is a lack of monitoring tools for quantifying the impact.
4) **Summer peaks period.** Thira main focus is however on the huge tourists flows arriving daily during summer. These flows are actually impacting the quality of life on the island with severe issues on traffic, water and electricity consumption as well as on waste management.
5) **Local products.** Thira particularly wants to promote local gastronomy, restaurants and wineries and local products.

The main question that Thira Municipality wants to answer is: How can we monitor the tourism flow and support the decision-making process? There is the need to have an up-to-date/near-real-time monitoring of the tourist arrivals/departures and movements. A very first list of questions – that will be further developed

during the project activities - are reported below. These questions will drive the data collection and analysis methods development:

a) Which is the origin Country of travelers? Thira needs to know the distribution of tourists by country of origin to better target marketing actions.
b) How long do the tourists stay in Santorini? Thira would like to monitor precisely the average length of the stay of tourists, possibly segmented for tourist category.
c) Are the first-time visitors and returning visitors behaving differently?
d) Can we predict the tourist arrival flows? There is a general interest for predicting the number of daily arrivals/departures from each of the few points of access to the island and the paths followed.
e) Which is the qualitative level of satisfaction / dissatisfaction of the user?

DATA

The datasets available at the Thira Municipality, or that can be collected by them or by the Consortium, are listed below. In the section, we will discuss the main characteristics of the data.

WHAT IS AVAILABLE NOW

In this section we report the findings about the datasets which are available now or can be provided in the future by the Thira Municipality.

FLIGHTS ARRIVALS AND DEPARTURES

From the Santorini airport we have the aggregated data of number of passenger arrivals and departures per month and per domestic / international flights. This dataset could be useful for understanding the country of origin of travelers. However, we have the origin of the flight but not the passenger nationalities and in many cases the origin of the flight is not the origin of the passengers. In addition, looking at the time series obtained from the air passengers dataset, we could extract the mobility patterns stratified by origin. However, in order to carry out this analysis, it is necessary that the air data contain the real origin of the journey and not only the last section directly connected to the destination airport. We will investigate the existence of this dataset and the possible process and cost to get these data.

This is historical data updated until two or three months before. We have started a communication with Civil Aviation Authority to have more detailed data (per day or per company), but so far, we received negative answers.

The datasets of all the flight data about Greece is publicly available at the URL:

- http://www.ypa.gr/en/profile/statistics/yearstatistics/
- http://www.ypa.gr/en/profile/statistics/temporarystatistics/

For each airport there is the aggregated number of arrivals and departures for domestic and international flights, as depicted in the snapshot of Figure 1.

CIVIL AVIATION AUTHORITY				SEPTEMBER 2018		
STATISTICS SECTION						
PROVISIONAL	COMMERCIAL TRAFFIC					
DATA						
	DOMESTIC			TOTAL INTERNATIONAL		
	FLIGHTS	PASSENGERS		FLIGHTS	PASSENGERS	
	ARR+DEP	ARRIV	DEPART	ARR+DEP	ARRIV	DEPART
PAROS	484	12851	15752	20	757	1018
RODOS	801	39658	42061	5155	402069	424297
SAMOS	402	7391	8783	408	25304	28700
SANTORINI	1520	77608	85979	1278	81306	92925
SITEIA	88	1191	1667	44	3288	3156
SKIATHOS	122	2699	3941	516	27952	38113
SKYROS	62	760	1098	10	294	346
SYROS	62	791	1312	0	0	0
CHANIA	361	26203	26867	2329	186459	193390
CHIOS	534	9511	12334	13	398	587
ATHENS	9230	436108	390361	11937	817120	906761
TOTAL	21870	926658	926114	45293	3197376	3538910

Figure 1. A snapshot of the Excel spreadsheet storing the data about the Hellenic Civil Aviation

From these tables we can extract the data relative to the Santorini Airport but also data relative to nearby islands like Mikonos or Crete that might be useful for comparisons. These excel files are also stored in the project repository.

More specific data about flight arrivals and departures, extracted from the official statistics, are available at the Santorini Airport web site
https://www.jtr-airport.gr/en/jtr/air-traffic-statistics

SANTORINI AIRPORT - 2018 vs 2017

Passengers	Domestic			International			Total		
Month	2018	2017	%Δ	2018	2017	%Δ	2018	2017	%Δ
JANUARY	30,390	27,931	8.8%	0	0		30,390	27,931	8.8%
FEBRUARY	30,459	29,811	2.2%	0	0		30,459	29,811	2.2%
MARCH	49,474	42,475	16.5%	4,171	763	446.7%	53,645	43,238	24.1%
APRIL	86,111	74,387	15.8%	40,067	26,059	53.8%	126,178	100,446	25.6%
MAY	134,794	113,786	18.5%	115,203	87,846	31.1%	249,997	201,632	24.0%
JUNE	155,362	132,225	17.5%	180,513	144,910	24.6%	335,875	277,135	21.2%
JULY	159,923	140,470	13.8%	231,729	206,952	12.0%	391,652	347,422	12.7%
AUGUST	155,053	133,662	16.0%	230,796	214,056	7.8%	385,849	347,718	11.0%
SEPTEMBER	163,789	135,898	20.5%	178,981	160,028	11.8%	342,770	295,926	15.8%
OCTOBER	126,392	110,368	14.5%	83,515	65,558	27.4%	209,907	175,926	19.3%
TOTAL JTR	1,091,747	941,013	16.0%	1,064,975	906,172	17.5%	2,156,722	1,847,185	16.8%

Figure 2. A snapshot of the table reporting the aggregated data about flight passengers travelling to and from Santorini Airport.

Specifically, we have monthly aggregated data distinct for domestic and international flights. In Figure 2 we see the comparison between 2018 and 2017 by month of the travelling passengers.
We also have data about the arrivals/departures per origin Country of flights (this is not informative about the origin Country of passengers) as reported in Figure 3 below for data about October 2018.

Santorini Airport *Greece*

Reporting Period **October 2018**

	Aircraft	Passengers		
Country	***Arr - Dep***	***Arriving***	***Departing***	***Transit***
Great Britain	138	9,338	11,805	-
France	122	7,639	9,102	-
Italy	132	6,158	8,546	442
Germany	98	4,908	6,734	862
Switzerland	42	2,519	3,132	-
Austria	36	1,929	2,538	-
Netherlands	20	1,161	1,786	-
Spain	18	866	1,183	-
Finland	8	308	637	-
Other Countries	*42*	*145*	*1,592*	*185*
Grand Total	**656**	**34,971**	**47,055**	**1,489**

Figure 3. Example of data reporting the arrivals and departures aggregated per origin Country of flights on October 2018.

SHIPS ARRIVALS AND DEPARTURES

Thira has access to the monthly aggregated data about the arrivals and departures of boats from the Port Authority. While the flight data provides information for tourists arriving from far countries, the data of the naval traffic can allow us to acquire information on the number of people who go to the island from the nearest countries or who have made a stopover on one of the surrounding islands. This data, in combination with others, can be useful for creating a flow forecasting model. It would be useful to have this data at the daily level, because, through the analysis of the differences between incoming and outcoming time series it might be possible to measure the length of stay. However, at the time of this deliverable is completed, we do not have access to a finer granularity than the month.
These tables distinguish between actual passengers, or cars/motorbikes or trucks as shown in Figure 4.

SHIPS 2017	Passengers		Cars		Motos		Trucks	
	Arrivals	Departures	Arrivals	Departures	Arrivals	Departures	Arrivals	Departures
January	11.223	4.805	1.988	537	167	116	578	460
February	9.022	7.135	1.255	670	220	166	706	627
March	16.66	10.45	2.446	785	572	266	987	770
April	47.01	39.665	3.394	1.711	881	466	1.145	937
May	61.047	63.199	1.623	1.205	487	486	662	1.424
June	104.25	119.208	2.407	1.902	709	742	1.268	1.232
July	141.884	167.713	3.824	2.946	1.112	1.114	1.358	1.304
August	138.59	173.706	3.742	4.417	1.877	1.802	892	1.074
September	115.105	139.15	2.219	2.829	663	1.105	615	904
October	70.599	75.146	1.571	3.082	404	1.023	879	1.061
November	13.259	16.805	1.317	2.882	213	642	727	762
December	7.268	11.603	800	2.52	102	269	635	660
Totals	**735.917**	**828.585**	**26.586**	**25.486**	**7.407**	**8.197**	**10.452**	**11.215**

Figure 4. Snapshot of the excel file containing the aggregated data from the Port Authority of Thira for 2017. There is a distinction between passengers and cargo and vehicles ships.

This dataset is open to the public.
http://www.santoriniports.gr/en/

CRUISE SHIPS BERTH ALLOCATION

Thira has the schedule of cruise boat arrivals and berth allocation.
This data is not public and its access is allowed only to consortium partners, therefore we do not report here a snapshot of the tables.
Essentially, this dataset contains the detailed data for each month, and each day of the month, the planned arrival time of cruise ships, the departure time, the estimated number of passengers. This has been done to organize a berth allocation to avoid the contemporary arrivals in the same day of more than 8000 passengers.
One idea could be to use these data to compute the differences between incoming and outcoming time series in order to try to measure the length of stay.
This dataset is available in the project repository under password.

AIS DATA

Thira has installed Automatic Identification System (AIS) receivers to have data about ships in the area. A historical dataset is available in the project repository under password. What we could do is to investigate if from these data we can approximately measure both the length of stay and the number of presences in the year.

WATER

We have water consumption data for village per quadrimester for the last few years.
This data is public and reports for each area of the island the water consumption for each category of users (e.g., shops, residentials, hotels). Figure 5 shows a snapshot of the data for the Akrotiri area in 2011.

01	AKROTIRI					
	Δ.Κ.	2011				
		MONTHS				TOTAL
		1-4	5-6	7-8	9-12	
01	HOUSES	2.115	2.126	2.905	2.682	9.828
02	SHOPS - NON HYGENE	0	0	0	0	0
03	HIGHLIGHT SHOPS	0	0	0	0	0
04	RESTAURANTS, COFFEE, BARS	44	94	114	97	349
05	HOTELS and APARTMENTS	60	229	238	74	601
06	ROOMS INCLUDED	85	109	344	215	753
07	PUBLIC SERVICES	45	50	10	25	130
08	PROFESSIONALS (?)	157	158	379	393	1.087
09	WAREHOUSE, CLOSED SPACES	0	0	0	0	0
10	AGRIGULTURE	0	0	0	0	0
		TOTAL				12.748

Figure 5. A snapshot of the water consumption data for the Akrotiri area in 2011.

This data can be useful if integrated with other punctual data about accommodations and tourist presence. The consumption figure could be used in combination with other data to create an attendance forecasting model.

WIFI

There is a public Wifi from the Municipality at several locations in the Island. The Wifi hotspot Map is depicted in Figure 6.

We have investigated whether it would be possible to get the (anonymized) access data from Wifi hotspot. This would give us a trace of the visits of the island, at least for those tourists who use the Wifi. The problem is that the Wifi service is in outsourcing to an internet provider. Thira municipality is investigating if their contract already includes these data. We have received some preliminary data in the form of aggregated plots for some of the main access points in the island, namely Thira, Oia, Thirasia and the rest of the island.

At the time of writing of this deliverable we are still waiting for more details on the data accessibility. In particular, we would like to know if it is possible to analyze some information related to the configuration of the mobile phones, such as the language. In this case we could think of using this data to calculate the origin of travelers. This data could also be useful for measuring the length of stay by considering the period in which the device, used by its owner, was registered on the island.

We cannot show a screenshot here of the plots since this data is not public. Everything is available in the project main repository.

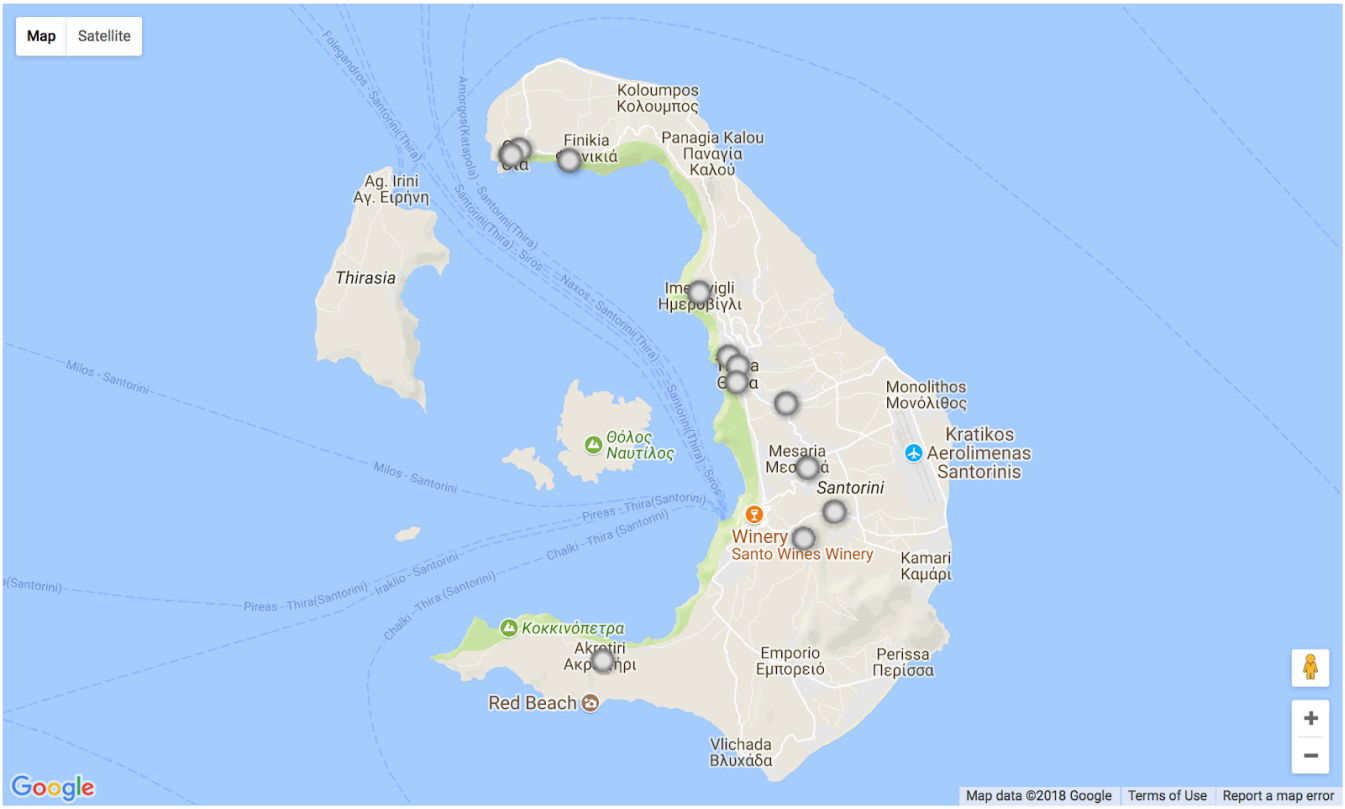

Figure 6. The map of the Wifi hotspots in Santorini island.

HOTEL PRESENCE

Thira municipality does not have the data about presence at the hotel. The city tax cannot be used as an estimation since this tax goes directly to the central government. Several documents have been given to the seconded researchers presenting the previous surveys and the results from the Tourism Observatory analysis during the last years. These documents (in Greek and translated with Google translate and with the help from Greek partners) are available in the project shared folder.

BUS AND TRANSPORTATION DATA

Thira municipality has no data about bus usage since the service is provided by a private company and the municipality has no means to force them to share their data.

ARCHEOLOGICAL SITES AND MUSEUM ENTRANCE TICKETS

Thira municipality has no data about the number of tickets sold by archeological sites since this is directly under the Ministry of Culture.

HOSPITAL AND EMERGENCY SERVICES

It would be interesting to be able to access the nationalities of people who access hospital services. In this way we could calculate the impact of tourists on health services and estimate the nationalities present in the area. We are investigating with the Thira Municipality if they have the possibility to get these anonymized and aggregated data.

OPENSTREETMAP DATA

The OpenStreetMap (OSM) project (www.openstreetmap.org) has collected an enormous amount of free spatial data about Santorini. The database contains geospatial features such as points, polylines and polygons mapping real-world elements such as roads, waterways, natural lands, points of interest. The dataset is available in shape file format at the link http://download.geofabrik.de/europe/greece.html.

WHAT THE CONSORTIUM CAN PROVIDE: SOCIAL MEDIA DATA

Santorini is a scenic location well covered in social media especially with photos and videos. We started a data collection from the official API of the most used media to better understand if this data can be used for MASTER activities. Photos are typically geotagged (e.g. Flickr) therefore they can represent paths followed by tourists. They also have several contextual data such as the content of the posts, other hashtags, follower/following friends and the user history of other locations visited. From this data we can roughly estimate the length of stay, too. A real time streaming of this data can also represent concentration of tourists at certain locations. This data can represent multiple aspect trajectories since the location is enriched with several contextual aspects. Social media data may allow the observation of the nationality of those presences, the frequency and the length of the visit. Applying some mining techniques, it is possible, in some cases, to be able to measure the sentiment of the post.

TWITTER STREAMING

We started a data collection from the Twitter API on May 8th 2018 and this collection is still ongoing. The collected tweets represent 1% of the tweets including the keyword "santorini" and having the bounding box of Santorini [25.310341,36.323836,25.511381,36.492272]. Twitter gives us tweets with bounding boxes within the indicated limits as well of tweets with bounding boxes **including** the indicated limits. E.g. A tweet which is geo-tagged with Greece bounding box will be filtered by our system, as it includes also Santorini.

By filtering tweets related to Santorini we can get a double perspective on the data:

- Content: Tweets including the keyword "santorini" indicate what is the content and context in which Santorini is mentioned and which users are speaking of Santorini. We wish to investigate whether these users mention Santorini before and after a possible visit.
- Geo-references: Geo-tagged tweets indicate the precise location where a tweet might be coming from. By collecting these tweets we hope to be able to recreate possible trajectories or, if the data is very sparse, get a glimpse of where most check-ins happen, a snapshot of where tweets are actually coming from.

Preliminary statistics on the data collection, by the end of 2018, are reported below:

Analysis Period	8.05.2018 - 31.12.2018
Total number of tweets	3.569.116
Tweets with coordinates	34.039 (0.866%)

Tweets with Place	362.016 (9.21%)
Tweets without geo-reference	3.535.077 (89.9%)

When looking deeper into the Tweets with Place, we analyze the type of Place (bounding box). We see that most of them represent the country (70%), then the city (15.1%) and administrative level (14.4%) while PoIs and neighbourhoods represent less than 1%.

We also look at the language the users are using and we see that most tweets are in English (36.3%), then Indian (16.4%) and Portuguese (15.5%) and other languages are less than 10%. Greek represents 5.02%, namely 179.116 tweets, of the whole collection.

FLICKR

We collected data from Flickr for 6 years (from 2012 to 2018) using the APIs provided by the platform.
The Flickr API allows us to collect retroactively the data even before the actual collection date and consequently before the project start. This is necessary in order to have enough data that allows us to perform meaningful statistics and trends over time. This is a common practice in the scientific fields of data science, machine learning and data mining. The idea is that the datasets are used to learn models of the past that can then be applied to more recent data. The pictures are collected using the function *flickr.photo.search* over a boundary box delimited by Santorini's coordinates: [25.310341, 36.323836, 25.511381, 36.492272]. The results are returned divided by pages, and the limit on the number of results per page is 500. To collect more data, we divided the box using a grid that moves from left to right and from top to bottom with step of 0.1 inside the Santorini box. We downloaded pictures for each sub-box moving from left to right and from top to bottom of the Santorini box. Some of the boxes in the grid may have a few pictures or none at all, but with this technique we are sure to collect as many results as possible.
Overall, we could gather 122,434 pictures. All of them are geotagged and have the time when the picture was taken by the user. In our dataset we have 1,556 distinct users and we collected some information about the users, as well.

Preliminary statistics on the data collection, by the end of 2018, are reported below:

Analysis Period	01.01.2012 – 31.12.2018
Total number of pictures	122.434
Pictures with coordinates	122.434 (100%)
Number of users	1.566

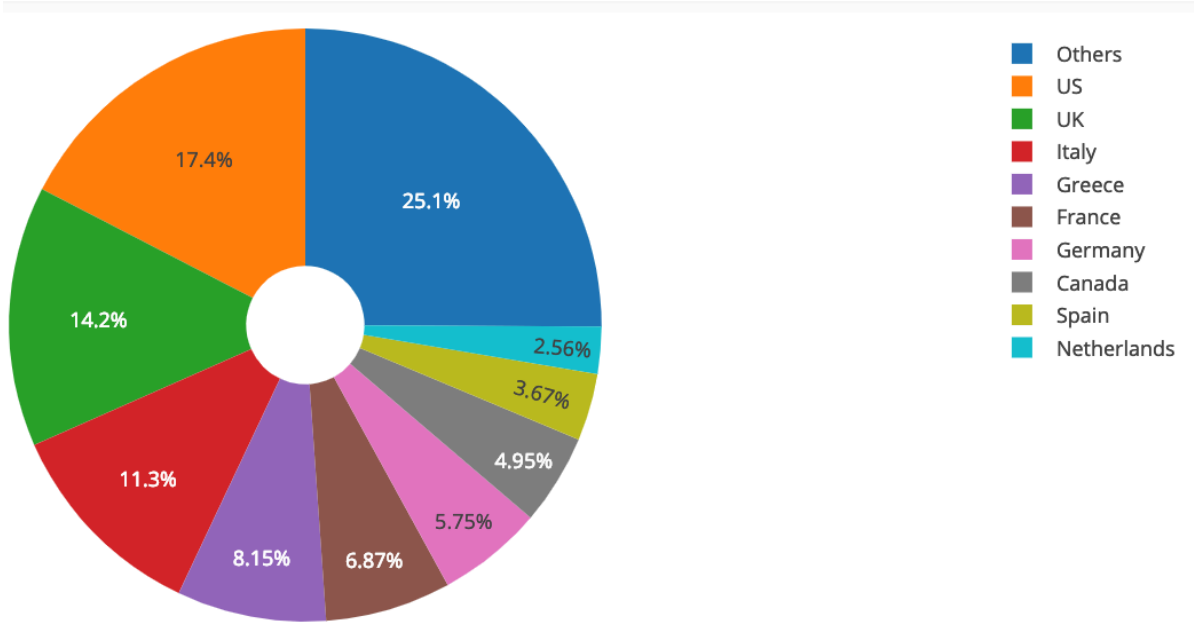

Figure 7. Distribution of user countries in Flickr's geotagged photos.

The location and timezone of the user can be used to conduct some analyses on their provenience. Statistics on the users' countries are shown in Figure 7.
Most of the users are from UK and USA, also a large percentage of Italians seems to visit Santorini followed by Greeks and French people. Although these results are biased by the usage of the Flickr platform (i.e., Flickr is more popular in some countries rather than others), they represent a good starting point for us since we can explore the trend behavior of tourists visiting Santorini.
In the future, we plan to investigate whether or not it is possible to collect more data for this platform.

The preliminary data gives us a general view and opens up several future possibilities. For example we want to establish if and how we can extract trajectories from the data and identify the starting and ending point of a tourist visiting Santorini. We want also to better understand if social media data can give a good estimation (in proportion) of the most crowded periods of the year and if we can identify the main topics of the discussions about Santorini.

HYPOTHESIS OF AD-HOC DATA COLLECTION

Unit Managers of Thira, CNR, UVSQ and UPRC in a meeting discussed about the hypothesis of running a data collection with BlueTooth LE placed in some crucial points of the island to monitor the flows of pedestrian or vehicles. In case of interest the consortium will discuss how the project might support this action, also in accordance with the Ethical Committee and Independent Ethical Advisor opinion. However, the conclusion of the discussion was that, after checking with stakeholders, it seems that the Municipality cannot proceed with this kind of action at the moment due to lack of personnel who can take care of the devices.

3. SEA MONITORING SCENARIO

This section of the deliverable is related to the activity of task 5.2, namely the Sea Monitoring application Scenario, whose responsibility is UPRC. The material of this part of the deliverable is mainly prepared during secondments to DAL, but also integrated during parallel activities of project partners.

The Secondments linked to the Tasks T5.2 are the following: HUA (Iraklis Varlamis & Konstantinos Tserpes) seconded to DAL in M5-M6, and UNIVE (Elisabetta Russo) seconded to DAL at M5, as presented in previous Table 1 for a total of 5,1 PMs.

INTRODUCTION

In the MASTER Technical Annex we have identified a sea monitoring use case regarding the design of techniques for the monitoring of the maritime domain, which is a key domain for the European economy (90% of the EU's external trade and 40% of its internal trade is transported by sea), while from a technical point of view it offers a typical scenario for the Big data challenges and requirements to be addressed. First, the growing number of sensors (in coastal and satellite networks) makes the sea one of the most challenging scenario to be effectively monitored. The need for methods able to scale in time and space the data processing of vessel motion data at sea is highly critical for maritime security and safety. Moreover, the data of these sensors, when fused with mobility data of vessels, produces an extremely interesting kind of holistic trajectories. The analysis of streaming data from multiple sensors is essential to detect as soon as they occur critical events at sea. This poses the emphasis on incremental clustering able to include new data into the data-at-rest already processed and on sequential methods able to detect critical events by continuously processing data. In addition, different types of data are available and only if properly combined and integrated these data can provide useful knowledge. Different sensor technologies are being developed and the data coming from multiple sources need to be cleaned up from inconsistencies, standardized in format and summarized. Finally, data measurements have an intrinsic uncertainty which proper fusion and clustering address to solve in the pre-processing phase (by assessing the quality of data themselves) and combining measurements from complementary sources. To address these issues, data-driven methods for the extraction and classification of the maritime patterns of life must be developed.

In MASTER, there are many academic participants with a strong enrolment to maritime domain. The present section discusses the requirements collected during the secondments of the involved partners.

GENERAL REQUIREMENTS

The following general observations and requirements emerged from the discussions during the involved secondments.

The recently revised approval of the European Union Maritime Security Strategy (EUMSS) - Action Plan, has further stated the importance of an actionable strategy towards a more focused reporting process to enhance awareness. Within this framework, discovering and characterizing the activities of vessels at sea are

key tasks to Maritime Situational Awareness (MSA) and constitute the indispensable basis for a fully capable Maritime Security (MS) plan. The achieved knowledge enhances the classification and prediction of vessel activities, as well as the detection of anomalous behaviors, enabling an effective and quick response to maritime threats and risks.

The recent build-up of terrestrial sensor networks and satellite constellations of Automatic Identification System (AIS) receivers is facilitating the availability of ship movement information, both in coastal areas and open waters. The resulting amount of information is increasingly overwhelming to human operators, requiring the aid of automatic processing to synthesize the behaviors of interest in a clear and effective way. Although AIS data are only legally required for larger vessels, their use is growing (e.g., their use has been recently extended to smaller fishing vessels according to the European Commission regulations in the Mediterranean Sea), and they can be effectively exploited to infer different levels of contextual information, from the characterization of ports and off-shore platforms to spatial and temporal distributions of routes. Machine learning and data mining techniques have emerged as prime candidates in order to analyze vessel traffic, although the approaches have to challenge their efficiency and effectiveness within the larger context of big data processing, analysis and visualization. Even the noise and the redundancy reduction as well as the fusion of complementary information become less trivial when processing spatio-temporal streams of big data. As the amount of the available AIS data grows to massive scales, any data mining and machine learning approach to vessel activity characterization must contend with the higher-level challenges posed by massive data analysis (e.g., data representation, sampling bias and inference), especially when these methodologies need to be validated operationally within security and defense constraints.

DATA-DRIVEN EXTRACTION AND CLASSIFICATION OF MARITIME PATTERNS OF LIFE

One category of requirements is to develop efficient methods/algorithms to analyze, process, integrate, compress and visualize heterogeneous big data streams in the maritime domain. Specifically, the challenge is to develop unsupervised spatio-temporal clustering techniques of vessel locations jointly with AIS static and kinematic information to extract a dictionary of historical vessel patterns-of-life. These vessel motion information layers can be combined with other contextual sources (e.g., environmental, seasonal, operational, political, economic conditions) which provide a refined integrated picture of the maritime activity in the area. The historical vessel patterns-of-life can be used to analyze the evolution of traffic in different time scales (weeks, seasons, years) in order to:

- provide support to the maritime spatial planning operations when new traffic separation schemes need to be assessed, new infrastructures at sea need to be built (e.g., wind off-shore platform, oil platforms);
- inform the traffic modeling and simulation techniques;
- help in refining Search and Rescue operations and identify vessels responsible for oil spill detections;
- enhance ship prediction.

Due to the intrinsic features of the maritime domain, the ship prediction task can be performed into three different ways, taking into account the different time scales. As an example, the short-term ship prediction provides track estimation on a near-term scale from the partially observed track. This kind of short-term forecast cannot go very long in the future since the length of the prediction window is affected by the growing uncertainty from the observed points, e.g. ships maneuver along the trajectory. Hence, for maritime

applications, a more relevant problem is ship route prediction. A route is a historical group of tracks that connect two points of interest (either ports or entry/exit gates). These routes (which are a motion model of the recurrent behaviors of vessels between two given points) can enhance a long-term forecast (in general up to several hours after the last observation). The classical available methods show generally lower spatial accuracy. A context-enhanced route prediction method can be employed whose accuracy in estimating future ship positions is improved by the exploitation of historical patterns of life. As part of the route prediction, ship destination prediction can be performed. This function will enable the validation of the field Destination in AIS messages. The main outcome of the data-driven route prediction is the enhancement of the route planning optimization, once traffic route forecasts are integrated with weather forecasts and environmental data in general. Specifically, historical patterns of life can help the route planning by integrating the derived historical routes and the METOC conditions (e.g., wave scatter diagrams, significant wave heights, wind distribution). This is for example the case when new traffic separation schemes need to be assessed, or new infrastructures at sea need to be built (e.g., off-shore wind farms, off-shore oil platforms).

DATA-DRIVEN EXTRACTION AND CLASSIFICATION OF MARITIME ANOMALIES

Additionally, the historical vessel patterns-of-life can be used as baseline prior information in order to detect critical events:

- Identifying possible stationary vessels based on speed gating algorithms (comparing the actual and the estimated future subsequent displacements of vessels).
- Identifying possible off-route vessels.
- Identifying possible high-speed or low-speed vessels.

The method detecting critical events can be validated by comparing them with some available reference points (regulations, protected areas, closed areas). A data-driven methodology to associate real-time tracks of vessels entering into a specific area to the derived traffic patterns of life. This work builds on unsupervised learning techniques which provide models for normal traffic behavior. This is done by the pre-extraction of contextual information as the baseline patterns of life (i.e., routes) in the area under investigation. The characterization and representation of the derived routes are developed in support of exploitable knowledge to classify anomalies. A hierarchical reasoning may be adopted where new tracks can first be associated with existing routes based on their positional information only and "off-route vessels" are detected. Then, for on-route vessels further anomalies can be detected such as "speed anomaly" or "heading anomaly".

The key assumption is: tracks that are found to sit in or close to one of these clusters may be considered normal tracks, while those that sit at a larger distance from all the clusters may indicate an anomaly. Following a hierarchical approach in the feature domain, the classification does not assume complete trajectories, since it is based on an incremental handling of the track, updating the classification as soon as a new contact from the same vessel is received. Consequently, the longer the track duration and/or the larger the number of feature measurements available, the better the track classification performance is expected, which is a key requirement. In addition, the explicit association of a vessel (or track) to an existing route is a requirement.

To meet the requirements, different scenarios will be used, such as: the use of patterns of life for vessel traffic monitoring to enhance route prediction; the use of patterns of life and predicted routes for the detection of anomalies at sea.

The trajectory data of vessels at sea is generally partially and sequentially observed, and the challenge relates to their classification and prediction as soon as new data feeds are received. In this setting, two main requirements are highlighted:

- Route plan improvement: The data driven route prediction algorithm can be used to detect better routes (e.g., in a specific day of the week);

- Enhanced anomaly detection: The scheme should address the following Maritime Situational Indicators of interest for operators: (a) "The vessel is off-route"; (b) "The vessel is in reverse traffic on the route"; (c) "The speed is not compatible with the route followed".

The analysis of routes by ship-type will provide also the possibility to test the anomaly indicator "The type of the vessel is not compatible with the route followed". The differentiation of route layers filtered by ship-type is considered also a requirement, in order to enhance the detection of these anomalies.

IMPROVE THE KNOWLEDGE OF THE FISHING ACTIVITIES IN THE NORTHERN AND CENTRAL ADRIATIC SEA.

The Northern and Central Adriatic Sea (Geographical Sub-Area GSA17), characterized by a wide continental shelf and eutrophic shallow waters, are well known to be intensively exploited. Indeed, the Adriatic fish stocks result overexploited due to the nonselective and unsustainable fisheries. In this context, the development of an effective fishery management plan is needed in order to achieve a sustainable exploitation of commercial species and guarantee a healthy, productive and resilient ecosystem. Different management measures are currently used in the Mediterranean Sea, and consequently in the Adriatic basin, among which: fishing limits, such as the control of fishing capacity (e.g. catch limits for the bluefin tuna, swordfish and recently also for anchovy and sardines in the Adriatic Sea); the reduction of fishing effort; the application of technical measures (mesh size regulation, minimum landing size for several important species), the establishment of closed areas (e.g. the permanent ban of trawling activities within 3 nm of the coast, established with the Council Regulation 1967/2006/EC[1]) and the closed season (e.g. the ban of trawling activities for a variable period, generally during the summer season). Moreover, in the Adriatic Sea, the fishing activities were periodically banned in the Jabuka/Pomo Pit, a key nursery ground, especially for hake and Norway lobster. In particular, demersal fishery was interdicted from July 26th, 2015 to July 26th, 2016 by the Italian Decree GU/7/2015[2] and currently the General Fisheries Council for the Mediterranean (GFCM) in the 41st session (Montenegro, October 2017) adopted the recommendations of the establishment of a Fisheries Restricted Area (FRA)[3] in the Pomo Pit, banning demersal fisheries. In order to assess the effectiveness of these measures and develop new ones, it is necessary to improve the knowledge of these activities and to monitor them in space and time.

In the last years, many efforts have been made to achieve these goals using useful tools, such as Vessel Monitoring System (VMS) and Automatic Identification System (AIS). Indeed, the knowledge of temporal and spatial variation and distribution of fishing activities is a key element for policy-makers and researchers. The VMS, a satellite-based fishing vessel monitoring system providing several data to the fishery authorities at

[1] https://publications.europa.eu/en/publication-detail/-/publication/f7b0a754-4a19-4cf9-8040-aeae2faace38/language-en
[2] http://95.110.157.84/gazzettaufficiale.biz/atti/2015/20150162/15A05454.htm
[3] Appendix 7 at http://www.fao.org/3/i8500en/I8500EN.pdf

regular intervals (about 1-2 hours), was introduced by the European Union, formerly for vessels length more than 15 m and from 1 January 2012 also for vessels above 12 m (1224/2009/EC)[4]. While the AIS, designed primarily as navigational aid to avoid vessel collisions, was introduced by the International Maritime Organization (IMO) with the International Convention for the Safety of Life at Sea (SOLAS), for ships with 300 or more gross tonnage (GT) and all passenger ships. Nevertheless, with the entry in force of the European Directive 2011/15/EU[5], also fishing vessels have the obligation to install the AIS device (from May 2012 all vessels with a length of more than 24 m overall, from May 2013 all vessels above 18 m and from May 2014 all vessels above 15m). Vessels transmit their position at variable rate, from 2 seconds up to two minutes. Despite the VMS were introduced exactly for the monitoring of fishing activities they have some limitation, such as long time between the transmission of two consecutive signals (low temporal resolution), as well as the difficulty to obtain the data. Hence, having the AIS data at a higher temporal resolution and being openly available to the public, the use of AIS data for scientific and managing purpose is growing increasingly. Nevertheless, also the AIS data have different disadvantages. The aim of the work was to improve the knowledge of the fishing activities in the Northern and Central Adriatic Sea, identifying the most exploited areas, the fishing grounds, as well as the annual and seasonal fishing behavior of the Adriatic's trawl fleet, and test the effectiveness of the current fishery managements and suggest possible implementation of new ones.

The main questions relative to the sea monitoring in the Adriatic Sea are:

- How can we improve the knowledge of the fishing activities in the Northern and Central Adriatic Sea?
- How can we evaluate the effectiveness of the current fishery managements?
- How can we detect the spatial distribution of commercial fishery catches?
- Is there a correlation between the fishing effort and the catches relative of that area? Is there a spatial pattern?
- Can we predict the distribution of the fishing effort and the relative catches?

At present, there is great scientific interest about the implementation of new approaches, namely the Ecosystem Based Approach, for managing renewable resources in marine environment. In particular, the improvement of our capability to map, in space and time, the real fishing effort and to assess effectiveness of the management strategies put in place are two of the most critical issues. Many efforts have been devoted to this, but there is the need for increasing exchanges among different labs and groups, at the international level, for integrating efforts and improve quality of expected outputs.

In order to answer some of these queries we want to use a Trajectory Data Warehouse (TDW), namely a data warehouse aimed at storing aggregate information on trajectories of moving objects, which also offers visual OLAP operations for data analysis. In our case, from the AIS raw data we have to reconstruct the trajectories of the fishing vessels and the data warehouse model includes both temporal and spatial dimensions: the hierarchy for the *spatial* dimension consists of a collection of regular grids of increasing size whereas for the *temporal* dimension, the hierarchy has as base granularity a one day interval. By using the data warehouse we are able to investigate fishing activities (in terms of both fishing efforts and catches) in the different areas of the Adriatic Sea and during different time periods, i.e., months, seasons, years. This is a great improvement with respect to the present "state-of-the-art", in which fishing effort and catches are not really spatialized. Moreover, all this allows us to obtain high resolution data to be compared with fishery-independent ones, which are instead based on annual samplings, as a consequence resulting in less informative maps compared

[4] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32009R1224&from=EN
[5] https://eur-lex.europa.eu/eli/dir/2011/15/oj/ita/pdf

with those obtained by the TDW. All this could be quite helpful in order to propose new management policies for reducing fishing pressure and improve capability to preserve renewable resources in marine environment.

DATA

WHAT IS AVAILABLE NOW

The datasets available at the UNIVE, DAL, HUA, or that can be collected by them or by the Consortium, are listed below.

SURVEILLANCE INFORMATION

Vessel position reports coming from multiple sensors and sensor-networks can be used to extract patterns of life. Among them AIS data (from coastal and satellite receivers) provides a vast amount of near-real-time information, calling for an ever-increasing degree of automation in transforming data into meaningful information to support operational decision makers. AIS is a self-reporting messaging system originally conceived for collision avoidance to broadcast information on their location (positional, identification and other information) at a variable refresh rate, which depends on their motion. AIS is mandatory for ships of 300 gross tonnage and upwards in international voyages, 500 and upwards for cargoes not in international waters and passenger vessels. In addition, fishing vessels greater than 15 m sailing in water under the jurisdiction of the European Union Member States shall also be required to be fitted with AIS. Vessels at anchor transmit their position every two minutes and increase the broadcast rate up to two seconds when maneuvering or sailing at high speed; every five minutes, vessels transmit other data (static and voyage related information) containing identifiers, such as International Maritime Organization (IMO) number, call sign, ship name and Maritime Mobile Service Identity (MMSI), used as a primary key to link the message to position information. Static information also includes size, type of vessel and cargo, whereas voyage related data, such as Estimated Time of Arrival (ETA) and destination, are manually set and not fully reliable. Over the last several years, the AIS data received by ships and coastal stations have been transmitted to regional or national data centers. When multiple receivers are connected into networks, certain challenges arise with data intermittency, resolving data redundancy received by multiple receivers, correcting errors in timestamps assigned by varying receivers and identifying tracks of vessels that erroneously share the message identifier. Receiving AIS messages from space is becoming increasingly commonplace. As opposed to terrestrial networks of AIS receivers, whose performance is characterized by high persistence, but limited coverage, satellite-based systems can pick up satellites, so the AIS coverage is global at the expense of persistence, due to the orbiting platform revisit time. It is clear that when integrating such systems with data received by terrestrial receivers, there are additional issues to resolve with variable frequency update, coverage and persistence.

Relevant information from AIS data are:

- Timestamp (reception time of the AIS frame in Unix epoch)
- MMSI_Number (the unique vessel ID)
- Longitude and Latitude (in WGS84 format)
- Speed over Ground (SOG)
- Course over Ground (COG)

•Ship Code (according to AIS specification; an 'ad hoc' routine has been developed to compute exact ship type from ship code)

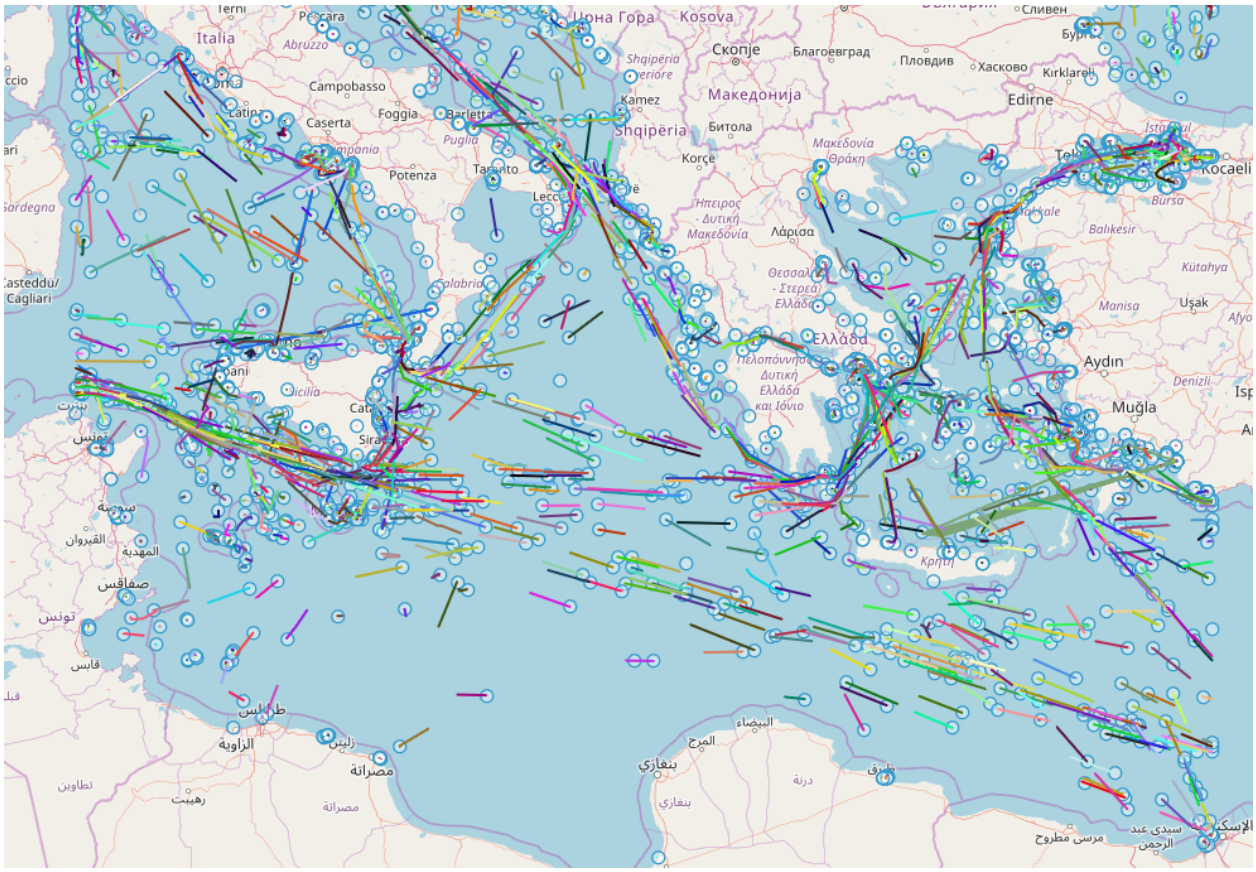

Figure 8. Instance of vessel trajectories in the Eastern Mediterranean

METEOROLOGICAL DATA

- Climate data is available free at the web site
  https://www.noaa.gov/
- Copernicus is the European Union's Earth Observation Programme offering information services based on satellite Earth Observation and in situ (non-space) data.
  https://www.copernicus.eu/en
- Some data about sea state conditions can be freely accessed from:
  http://www.ionioproject.eu/
  http://www.sea-conditions.com/en/
  The dataset comprises a set of parameters for characterizing the environment including the Significant Wave Height (SWH).
- Mediterranean Sea Physics Analysis And Forecast Data
  http://www.myocean.eu/web/69-myocean-interactivecatalogue.
  php?option=com_csw&task=results&simplesearch=ok&advancedsearch-
  geographical_area[]=advancedsearchgeographical_area-mediterranean-sea
- Contextual Information
  Maritime regulations; Traffic separation schemes
  Maritime protected areas (whose locations are visualized as kml files):
  earth.google.com/gallery/kmz/marine_protected_areas.kmz

http://atoll.floridamarine.org/Quickmaps/KMZ_download-bounds.htm
Closed areas, Anchoring areas

AIS DATASET FOR THE NORTHERN AND CENTRAL ADRIATIC SEA (GSA17)

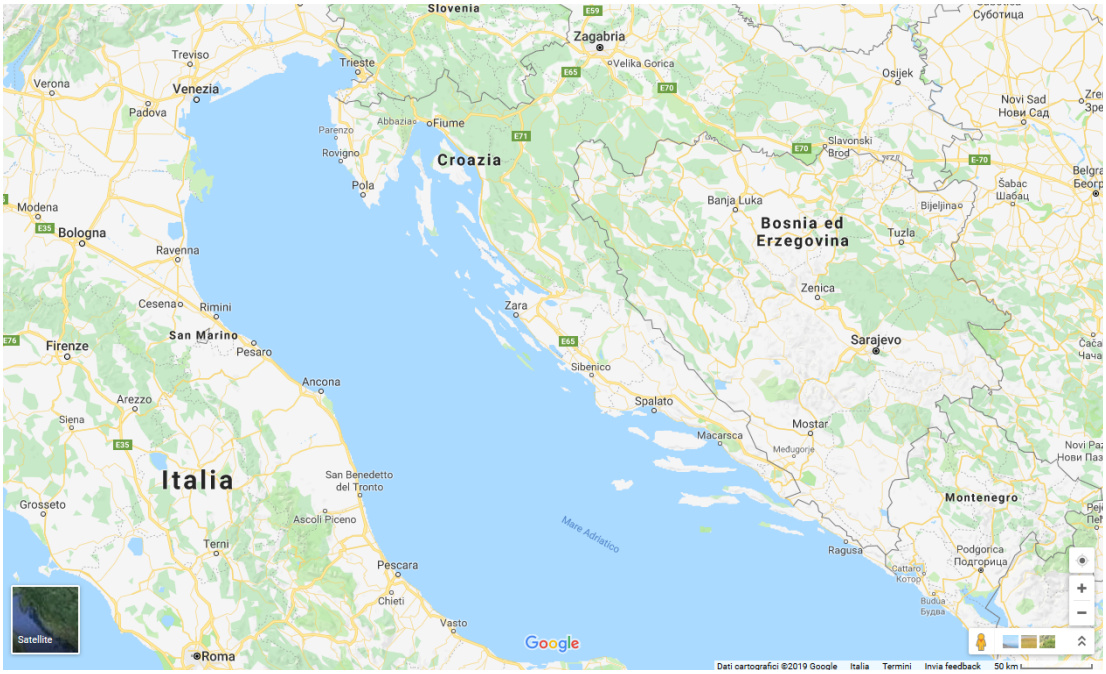

Figure 9. Map of the northern and central Adriatic Sea (GSA17)

The area of interest is the northern and central Adriatic Sea (FAO Major Fishing Area 37.2.1; FAO Geographical Sub-Area [GSA] 17), a sub-basin of the central Mediterranean Sea depicted in Figure 9. The GSA17 is well known to be the most productive area of the entire Mediterranean Sea and one of the most exploited of the European Seas.

The northern basin is characterized by an extended continental shelf (average depth of 35 m), the widest in the Mediterranean Sea, and eutrophic shallow waters, while the central part is deeper, arriving at 270 m of depth in the Pomo/Jabuka Pit. The water circulation is cyclonic and driven by the rivers discharge, mainly comes from the Po River, and the surface wind stress (such as the Bora).

Globally, the studied area covered 74,965 km^2 including three Economic Exclusive Zones (EEZs): 39,946 km^2 of Croatian waters; 34,806 km^2 Italians and 213 km^2 Slovenians.

The AIS raw data has been provided by the Italian Coast Guard (ITC and Traffic Monitoring Department – Rome) and consists of around 922 million signal positions released by Adriatic fishing vessels operating in the northern and central Adriatic Sea (GSA17) during the period between January 2015 and December 2016. Data of position (latitude and longitude), speed, time (Unix time) and the MMSI number are used to analyze fishing activities of 639 and 643 vessels for the 2015 and 2016, respectively. In order to identify each fishing vessel, the call sign (IRCS in the Fleet's Register), the ship's name and the MMSI number are used. First, the call sign and the ship's name, reported in the AIS data, are linked to the EU Fishing Fleet's Register

(http://ec.europa.eu/fisheries/fleet/index.cfm), which provides information about the length overall (LOA), gross tonnage (GT), primary and secondary gears. Nevertheless, in some case (e.g. erroneous call sign or ship's name) the Marine Traffic website (https://www.marinetraffic.com/) is used for the vessel identification using the MMSI. Moreover, in order to overcome misclassification problems due to errors occurred in the Fleet Register, the speed frequency distribution is used to a more accurate identification of the fishing gears.

FISHING GEARS

The analysis of the fishing activities is related only to actively towed fishing gears (trawling), and in particular to small and large otter trawls (SOTB and LOTB, respectively), Rapido, which is a particular kind of beam trawl (RAP), and midwater pair trawl (PTM, called also Volante). These gears represent the largest portion of the Adriatic fishing fleet above 15 m of length. The Adriatic trawl fleet was represented by Italian (~ 90%), Croatian (~ 10%) and Slovenian (<1%) vessels.

LANDING DATA

Landing data were obtained from the Chioggia's Fish Market, whose harbor hosts the major fishery fleet of the Adriatic Sea. The landing dataset included 104 commercial species caught during the biennium 2015-2016 in the northern Adriatic Sea (GSA17). These data were associated with the trajectories of each fishing vessel, and maps of catches distribution were produced.

4. TRANSPORTATION SCENARIO

This Chapter of the deliverable is related to the activity of collecting requirements and datasets relative of Task T5.3, namely the Transportation Scenario, whose responsible is UNIVE. The material of this part of the deliverable is mainly prepared during secondments to PUC, but also integrated during parallel activities of project partner UNIVE.

The Secondments linked to the Tasks T5.3 are reported in Table 1 at the Introduction section of this document and are the following: UVSQ (Karine Zeitouni) at PUC at M6 and CNR (Vinícius Monteiro de Lira) at PUC in M10 for a total of 1,26 PM.

INTRODUCTION

The objective of Task 5.1 is to **study methods** for:

1. creating a public transportation observatory for buses data;
2. improving traffic prediction;
3. improving ride sharing methods to reduce the private vehicle usage.

By public transportation we refer to the forms of travel offered locally that enables people to travel along designated routes. Typical examples of forms of public transportation include buses, waterbuses, trains, trams and bike-sharing. We will exploit the data from PUC and UNIVE and datasets collected from social media in order to develop a prototype application.

In this deliverable, we concentrate our effort in collecting application requirements and datasets that could be used in studying methods for supporting the three objectives listed above. It is important to observe that it is out of the scope of the MASTER project to create a full transportation observatory, which is clearly an activity of the Transportation Company or a local Municipality. The goal of the project is to collaborate with PUC and UNIVE researchers, who are in turn collaborating with their local transportation companies (e.g. Bus Rio and ACTV in Venice), in collecting requirements and datasets that MASTER researchers can use for developing research methods that can support the improvement of public transportation services through a transportation observatory. One of the goals of a **transportation observatory** is to assess comprehensively how the public transportation systems of the city of Rio de Janeiro can effectively provide means to the inhabitants to reach the different *functional areas* of the city (e.g. tourism, shopping, and hospital areas). Rio de Janeiro, with an area of 1.255 km^2, is one the largest cities in Brazil with more than 6 million inhabitants and more than 6 million international tourists per year. For the case of the city of Venice, partner UNIVE, thanks to the collaboration with the local transport company ACTV, highlighted the importance of understanding the flows of people, because there is a strong need of distributing tourists in time and space. In fact, during holidays, in Summer and at Carnival, the city is overcrowded and this makes it hard for residents to execute daily activities. A worrying phenomenon is the constant increase of tourists whereas the number of residents is decreasing. One of the measures under analysis by the local municipality is to establish a fixed number of tourists according to the different periods of the year, in order to limit the number of people entering the city and avoid the city to be overcrowded. Hence, the transportation observatory can be a useful means to analyse the flows and detect residents and tourists different behaviour.

As for point 2) **improving traffic prediction**, thanks to the collaboration of partner UNIVE with the local transportation company ACTV, the consortium can access the ticket stamps and this will allow to develop methods to predict the flows of people in different periods of the year. This will be helpful for ACTV to devise a correct planning of waterbuses and boats in order to offer a satisfactory service and suitably arrange boat maintenance. Note that boat maintenance requires a boat stop for several months and this can affect the quality of service. Hence it is crucial to know how the boats request varies along the year. Moreover, for the municipality of Venice it would be very important to predict tourist flows in order to adequately establish the number of people allowed to enter the city.

As for point 3) **improving ride sharing methods to reduce the private vehicle usage** we have the possibility of analyzing the Bike Rio dataset presented below to better understand how to improve the bike sharing service and therefore reduce vehicle usage through the use of sharing bikes. Further discussions with PUC researchers, who in turn collaborate with the Rio de Janeiro transportation services, are needed to better focus the specific requirements and the methods that MASTER can provide. This will be better detailed in the next version of the Deliverable D5.2 at M24.

Concerning Social Media, PUC can provide a historical data collection of tweets related to traffic events. However, the dataset available at PUC has a quite old temporal period since it covers the year 2014, therefore we are not sure if and how this dataset can still be useful for building traffic models. During the secondment to PUC we did not go more in deep about this dataset and we leave to future secondments and discussions with PUC researchers to take a decision about the usefulness of this dataset and if the collection of new and more recent Social Media datasets can be really useful for the project.

GENERAL REQUIREMENTS

PUBLIC TRANSPORTATION IN RIO DE JANEIRO (BRAZIL)

During secondments the MASTER secondees interacted with the PUC researchers, who in turn are collaborating with the Municipality of Rio de Janeiro and Transportation services. They found that a need, for the final objective of studying methods for supporting a public transportation observatory for buses data, is to study the coverage of the bus and metro systems to the different function areas of the city (e.g. tourism, shopping, and hospital areas). PUC partner has a valuable dataset for this purpose. The dataset contains the schedules and itineraries of the circulating buses and metro railways of the city. One requirement that emerged is to study the transportation movements towards the different neighborhoods of the city taking into account the semantics involved in the areas of interest, the spatial location of these areas, the geographic of the neighborhoods, and the dynamicity of these areas according to the time of the day.

PUBLIC TRANSPORTATION IN VENICE (ITALY)

Venice is located between the Adriatic Sea and the Po' Valley and built on an archipelago of 117 islands in a shallow lagoon served by 177 canals with land areas connected by 409 bridges. The city is divided into the historical centre, the mainland urban settlement of Mestre (270,000 inhabitants) and the industrial district of Porto Marghera. The historical center is the largest urban car-free area of Europe, with passengers, cars and

heavy vehicles moved by boats and ferries on the larger canals. The main passenger modes are motorized waterbuses (*vaporetti*) and private taxis, which cover regular routes along the major canals and between the city's islands, transport boats (*moto-topo*) and private boats.

ACTV is the main public transport company that counts 151 waterbuses. It carries 145 million passengers a year on the Navigation network. It has more than 120 floating stations (jetties) and 27 well-connected lines. The bus network consists of 95 routes and the fleet is composed by 568 buses and 20 trams. The number of passengers is around 70 million a year.

ACTV started a collaboration with Ca' Foscari University of Venice in order to study the flows and to understand the behavior of its users. Some of the questions of interest are the following:

1. How do the stamps of the users vary in space and time?
2. Is it possible to classify the users into categories, like workers, students, tourists?
3. Are there typical patterns in the movements of the users?
4. Are there different behaviors during the weekday and at the weekend?
5. Which are the common itineraries for tourists?
6. Emphasize the differences between workers and tourists with respect to the most used stops, the time period and the duration of the stay in the town.
7. Detect the behavior of tourists during the different days of their stay.

DATA

TRANSPORTATION IN RIO DE JANEIRO

Here we report the datasets related to transportation in Rio de Janeiro, which belong to PUC partner and are accessible only on-site.

BIKE-RIO

This dataset contains bike trajectories from a public bike-sharing system used in the city of Rio de Janeiro called: Bike Itaú (https://bikeitau.com.br/). The system started on February 20th 2018, and is sponsored by the municipal government of Rio de Janeiro in partnership with a private Brazilian bank. The bike sharing system counts 2600 bicycles available at 26 stations located throughout several neighborhoods in the city.

BUSES-RIO

The dataset has past trajectories of the lines of buses circulating in the city of Rio de Janeiro. The trajectories were retrieved by PUC from the open data portal of urban mobility of Rio de Janeiro, available at [http://data.rio/dataset/pontos-dos-percursos-de-onibus], which is provided by the City Hall. Data from GPS of buses that operate in the city is continuously captured, in about every 1min 30sec. Each entry contains a timestamp, the bus identifier, the line number, the position (as latitude and longitude) and the speed, as illustrated in Figure N. 10. However, the public data portal of City Hall only provides the instantaneous data, i.e. no historical data is available and the data has to be collected explicitly with a streaming application. The historical dataset currently contains more than 3 billion samples in CSV format. To support experiments in developed methods, thus PUC previously collected historical samples from

June 12th, 2014 until February 28th, 2017, which represent almost 3 years of bus trajectory data. These datasets are available for use to secondees.

Timestamp	Bus_id	Line	Latitude	Longitude	Speed
13-09-2015 00:00:01	C27109	940	-22.827141	-43.294739	32.0
13-09-2015 00:00:01	C41401	303	-22.857653	-43.245167	0.7
13-09-2015 00:00:01	C50112	301	-22.929371	-43.253754	0.0
13-09-2015 00:00:01	C51512	738	-22.877365	-43.368198	0.0
13-09-2015 00:00:01	C72081	805	-22.889046	-43.292263	0.2
13-09-2015 00:00:01	C82596	363	-22.858412	-43.371071	0.9
13-09-2015 00:00:01	D58684	840	-22.841305	-43.371494	2.8
13-09-2015 00:00:01	C72081	805	-22.889046	-43.292263	0.2
13-09-2015 00:00:01	C82596	363	-22.858412	-43.371071	0.9
13-09-2015 00:00:01	D58684	840	-22.841305	-43.371494	2.8

Figure 10 - Snapshot of GPS observations data collected from PUC using the data.rio streaming API

POINTS OF INTEREST OF THE CITY OF RIO DE JANEIRO

We have built a dataset with the points of interest (POIs) within Rio de Janeiro. We have collected the data using the Yelps API (https://www.yelp.com/developers/). We used the geometry of the city to guide our crawler to collect POIs from the different points of the city. A snapshot of the collected data is shown in Table 2 below. Currently, the dataset contains 24.496 POIs. This dataset is available at the project repository under password.

Table 2 - A snapshot of the POI collected in Rio de Janeiro (JSON format)

{"query": {"latitude": -22.90088896885463, "longitude": -43.74014698536609, "radius": 2000}, "response": {"businesses": [], "total": 0, "region": {"center": {"longitude": -43.74014698536609, "latitude": -22.90088896885463}}}}
{"query": {"latitude": -22.95248091008819, "longitude": -43.528514458721524, "radius": 2000}, "response": {"businesses": [], "total": 0, "region": {"center": {"longitude": -43.528514458721524, "latitude": -22.95248091008819}}}}

TRANSPORTATION IN VENICE

Here we report the datasets related to public transportation in Venice, which have been provided by ACTV to UNIVE and are accessible only on-site. ACTV keeps the ownership of all these datasets. The datasets are the following:

1. User Profiles;
2. Residence data:
3. Stamp datasets.

Moreover, for our analyses we collected also **General Transit Feed Specification** (**GTFS**) about waterbuses and buses in Venice.
The detailed description of these datasets is given in the next sections.

USER PROFILES

The file contains the description of the user profiles who hold the card "VENEZIA UNICA". As shown in Table 3, it consists of two fields:

1. Profile_Code: a unique code profile.
2. Description: a description of the type of user.

Profile_Code	Description
138	138 - DIPENDENTI Actv/Vela
140	140 – LIDO Ordinario
125	125 - Solo BUS Studente
75	75 - Comune VE fasce deboli 20%

Table 3 - Profile codes for the "Venezia Unica" dataset

The field Profile_Code is used to establish the different ticket fares and it is based on the residence, particular health condition, ACTV employee and type of service (e.g., only bus). It consists of 168 records. It is worth noting that this dataset contains only categories of users which do not refer to specific individuals.

RESIDENCE DATA

The dataset contains information about the holders of "VENEZIA UNICA". The dataset is anonymized due to privacy reason. Each record has six fields:

1. ID_Client: a unique random number
2. Profile_Code: the profile code of the user
3. Description: a description of the user profile.
4. Address: address of the user
5. ZIP: ZIP code
6. Town: Residence town
7. Prov: Province
8. Location: A reference about the location of the user.

The dataset contains 65000 users and it includes people living not only in the islands but all over the world.

We filtered, cleaned, corrected and georeferenced this dataset and we extracted a new one, storing only users living in the main island. The resulting number of records is 9468. We cannot show an example of this data because it is not public. Hence, we present only a heat map of the geolocalization of the addresses in Figure 11 from which we can see the parts of the city with more users (darker red). We plan to use the

residence information to make clusters of users and to partition Venice in several areas in order to build an origin-destination matrix. This kind of matrix is crucial to understand flows of residents.

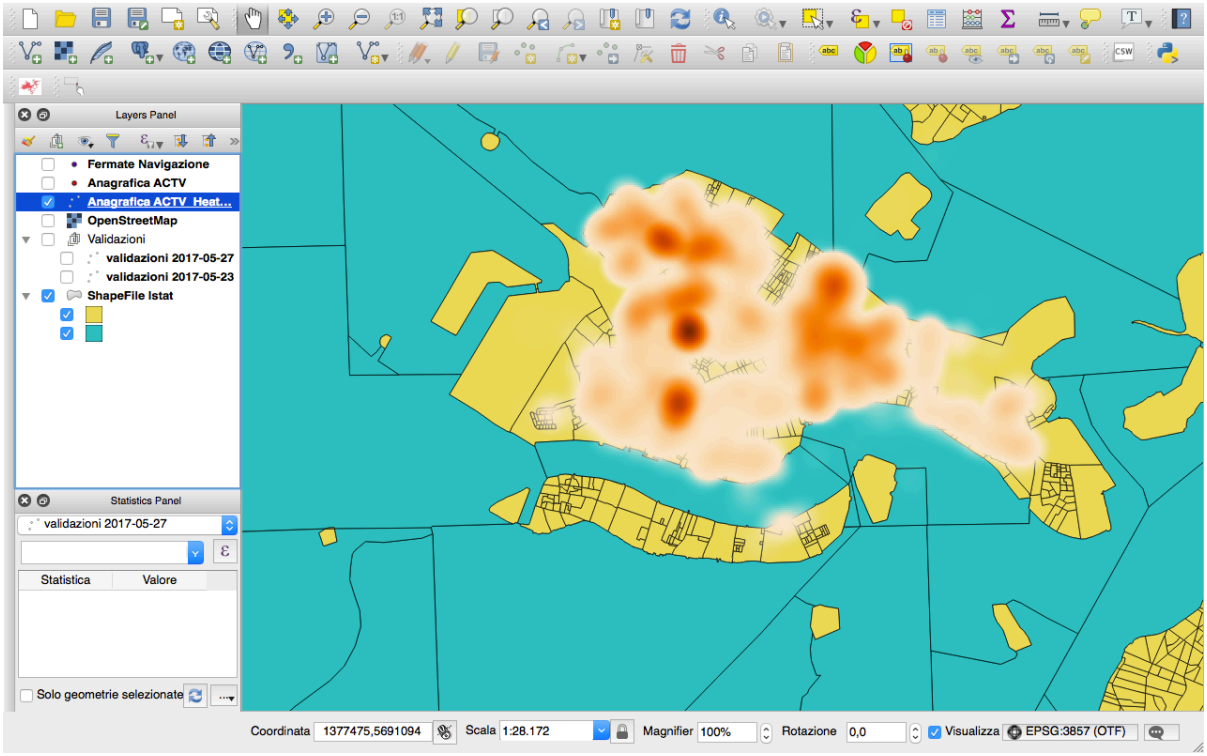

Figure 11. Screenshot of an heatmap of ACTV residence data

STAMP DATASETS

ACTV provided us two datasets containing the stamps of different kinds of tickets. The common format of these datasets is the following:

1. Date and hour of the stamp
2. Serial number of the user
3. Code_Profile
4. Code of the stop in which the ticket has been stamped
5. Name of the stop.

The Serial number of the user does not correspond to the ID_Client stored into the Residence Dataset. This is due to privacy reason. ACTV replaced the ID_Client with a random number keeping the same number for the same user if s/he stamped more than one time in the period of interest.

The first dataset contains 441000 stamps from January 15 2018 to January 21 2018 made by users holding a monthly or yearly ticket. Figure 12 illustrates how the data are spread over the week.

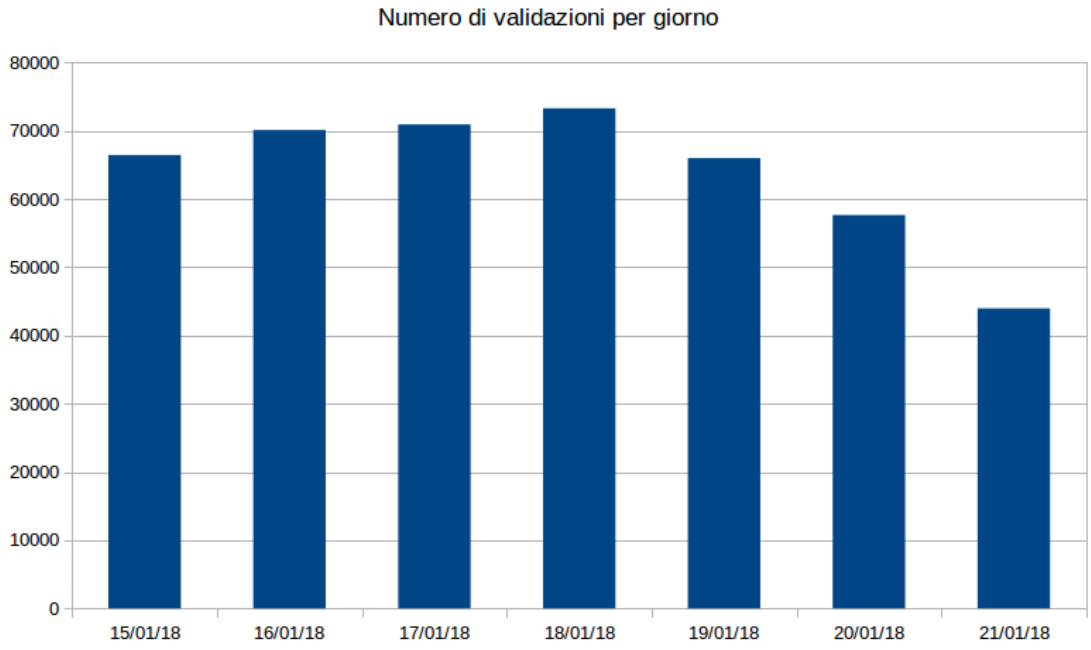

Figure 12. A snapshot of a plot showing the number of stamps per day spread over the week

The second dataset consists of 4876778 records from September 11 2018 to November 12 2018. It includes the stamps of all kinds of tickets. The dataset has been cleaned by removing the duplicates and we selected only the **time limited tickets.** Such tickets allow an unlimited number of trips and can be used on all transport means, both navigation and Mestre Urban Network and Lido buses involved in Venice Municipal area urban routes. They can have four different durations: 1 day, 2 days, 3 days or 7 days. They are usually bought by tourists. The resulting dataset consists of 3175730 records and the number of distinct users is 551962. Moreover, they are distributed in the following way among the different durations:

Table 4 - Numbers from the Stamps dataset of ACTV Venice

Duration of the ticket	Number of Stamps	Number of distinct users
One day	1211444	322732
Two days	661925	103546
Three days	834084	99957
Four days	350348	25727

These are preliminary datasets to experiment our techniques and we plan to ask ACTV other datasets for different time periods to verify the validity of our methods.

GENERAL TRANSIT FEED SPECIFICATION

The **General Transit Feed Specification** (**GTFS**) defines a common format for public transportation schedules and associated geographic information. It is a .zip archive containing the following files:

1. agency.txt: information about the transit agency such as, name, website and contact information.
2. calendar.txt: service patterns that operate recurrently like every weekday.

3. routes.txt: name and type of a route
4. trips.txt: list of the trips provided within a certain route.
5. stops.txt: geographic locations of each stop
6. stop times.txt: for each stop in a trip, arrival and departure time

The company ACTV provides and updates GTFS data for the navigation and bus networks through the web page http://actv.avmspa.it/en/content/catalogo-dei-dati-metadati-e-banche-dati-actv which can be accessed freely.

CONCLUSIONS

This deliverable reports "*the results of the requirement analysis for the scenarios and details application questions and data available*". As reported in the MASTER Grant Agreement Annex 1 Part A and B we have identified three application scenarios linked to three relative tasks: T5.1 tourism, T5.2 sea monitoring and T5.3 transportation. In this deliverable, we have reported the application requirements, questions and datasets collected so far for the three scenarios.

The content of the deliverable has been produced during secondments linked to WP5 and parallel activities of project partners. The total effort in PM for secondments linked to WP5 is 9,97PM. The 3,61 PM out of 9,97 are linked to T5.1 Tourism, 5,1 PM are linked to Task 5.2 Sea Monitoring, while 1,26 PM are linked to Task 5.3 Transportation Scenario.

As for Task 5.1 (i.e. tourism scenario) the main results in terms of application requirements is to study methods to monitor the tourism flow and support the decision-making process in the Santorini Island as a requirement by the Municipality of Thira. This is translated to different more specific questions, for example "Which is the origin Country of travelers?", "How long the tourists stay in Santorini?" or "Can we predict the tourist arrival flows?" to name but a few. The datasets available at Thira and made available to the consortium are several, such as the flights' and ships' arrivals/departures, water consumptions, reported in Chapter 2. The consortium also provided additional data by collecting historical datasets from Social Media.

As for Task 5.2. (i.e. sea monitoring scenario) one requirement is about the knowledge of the fishing activities in the Northern and Central Adriatic Sea, identifying the most exploited areas and the fishing grounds. A dataset with Adriatic sea fishing boats has been provided by UNIVE. From the secondments to partner DAL it has emerged important requirements for sea monitoring, such as, to identify possible stationary vessels based on speed gating algorithms and possible off route vessels.

As for Task 5.3 (i.e. transportation scenario) partners PUC and UNIVE provide to the consortium some historical datasets of buses and bikes in the cities of Rio de Janeiro and Venice. As requirements, we collected the need to study the reachability of public transportation in different areas of the city of Rio de Janeiro and the need to study the tourists flows in the city of Venice. As we already discussed above, the Social Media data that PUC can provide is quite old and we believe this cannot help, as it is, in developing a prototype application that is the reason why we did not include it in the current deliverable. We will consider, in the revised version of the Deliverable that is D5.2, whether it could be useful to collect other social media data or if using alternative datasets can still be sufficient to accomplish the Task objectives.

As we discussed above, the role of the consortium is not to create a public transportation observatory that is the task of the appropriate municipalities and transportation companies, but, as researchers, is to develop methods that can be used to create such transportation observatory for buses data. For the objective of "improving traffic prediction" we already discussed above how the Venice datasets, thanks to the collaboration of partner UNIVE with the local transportation company ACTV, will allow the Consortium to develop methods to predict the flows of people in different periods of the year. As for the objective of "improving ride sharing methods to reduce the private vehicle usage in comparison with other actors in the city" as we already discussed above, we plan to study the bike sharing datasets and improve the use of bike sharing thus reducing the use of private vehicles.

The applications that we plan to derive from the data analysed in this deliverable are related to the respective tasks, their datasets and collected requirements (i.e. tourism, sea monitoring, transportation) and they will be identified in Deliverable D5.2 due at M24 when the datasets and requirements will be revised, consolidated and selected for the prototype applications.

We have indeed planned a second version of this deliverable, D5.2, collecting a revision of requirements and datasets due at M24 with the objective of selecting which datasets and which requirements will be realized in the prototypes. The current deliverable is therefore an input to D5.2 that in turn will be an input to D5.3 at M36. The deliverable D5.3 will report about the research prototypes to be developed starting from the requirements and the datasets collected that in turn will be an input to D5.4 due at M48. This last deliverable will describe the developed prototypes and the experiences conducted in the diverse application scenarios.