D4.3 OPEN DATASETS FOR 6G-NTN DATA DRIVEN RADIO ACCESS NETWORKS

Revision: v.1.0

Work package	WP 4
Task	Task 2
Due date	30/08/2023
Submission date	29/09/2023
Deliverable lead	CTTC
Version	1.0
Authors	Husnain Shahid (CTTC), Miguel Ángel Vázquez (CTTC), Fanny Parzysz (OR), Laurent Reynaud (OR).
Reviewers	QCOM, UNIBO
Abstract	This deliverable identifies the open datasets as an initial input to design the data driven AI radio intelligent controller in order to effectively manage the resources of multi-frequency multi-constellation system. Moreover, it also intends to incorporate the heterogeneous properties of user terminal traffic for the identified use cases in D2.1. In this context, this document provides the guidelines about the attributes and the utilization of the identified open datasets for 6G-NTN project. This aims further to be used in designing the Deep Learning based radio resource management solutions.

www.6g-ntn.eu

Grant Agreement No.: 101096479
Call: HORIZON-JU-SNS-2022

Topic: HORIZON-JU-SNS-2022-STREAM-B-01-03
Type of action: HORIZON-JU-RIA

Keywords	Open-Source Datasets, Radio Intelligent Controller, Resource Management, Artificial Intelligence.

Document Revision History

Version	Date	Description of change	List of contributor(s)
V0.1	25/05/2023	Proposed table of contents.	H. Shahid, M. Á. Vázquez
V0.2	16/08/2023	Including comments from Orange and final inputs from H. Shahid.	H. Shahid, M. Á. Vázquez
V0.3	18/09/2023	Consolidating reviews from QCOM and UNIBO.	M. Á. Vázquez.
V1.0	28/09/2023	Final version ready for the submission to EC	CTTC, ORA, UNIBO

DISCLAIMER

Co-funded by the European Union

6G-NTN (*6G Non Terrestrial Network*) project has received funding from the **Smart Networks and Services Joint Undertaking (SNS JU)** under the European Union's **Horizon Europe research and innovation programme** under Grant Agreement No 101096479.

COPYRIGHT NOTICE

Project co-funded by the European Commission in the Horizon Europe Programme		
Nature of the deliverable:	**DATA**	
Dissemination Level		
PU	*Public, fully open, e.g., web (Deliverables flagged as public will be automatically published in CORDIS project's page)*	✓
SEN	*Sensitive, limited under the conditions of the Grant Agreement*	
Classified R-UE/ EU-R	*EU RESTRICTED under the Commission Decision No2015/ 444*	
Classified C-UE/ EU-C	*EU CONFIDENTIAL under the Commission Decision No2015/ 444*	

Classified S-UE/ EU-S	*EU SECRET under the Commission Decision No2015/ 444*	

* R: Document, report (excluding the periodic and final reports)

DEM: Demonstrator, pilot, prototype, plan designs

DEC: Websites, patents filing, press & media actions, videos, etc.

DATA: Data sets, microdata, etc.

DMP: Data management plan

ETHICS: Deliverables related to ethics issues.

SECURITY: Deliverables related to security issues

OTHER: Software, technical diagram, algorithms, models, etc.

EXECUTIVE SUMMARY

Deliverable D4.3 mainly contributes to identify the open-source datasets for designing the Artificial Intelligence (AI)-enabled Radio Intelligent Controller (RIC) to efficiently manage the resources of multi frequency multi-constellation Non-Terrestrial Networks (NTN) in the 6G-NTN end-to-end system. By effectively implementing the strategies for managing the resource allocation, the capabilities and reliability of NTN can be enhanced. This leads to effective spectrum utilization while minimizing the unmet system capacity, meeting the Quality of Service (QoS) requirements, and enhancing the overall system optimization.

In turn, a wide range of applications and services in various domains can be enabled. In this regard, D4.3 takes as input some of the Use Cases (UCs) defined in D2.1 "*Use case report*". In particular, the selected UCs are those that can be supported by RIC and include Maritime Coverage for search and rescue coast guard intervention, Adaptation to Public Protection and Disaster Relief (PPDR) or Temporary Events, and Consumer Handheld Connectivity and Positioning in Remote Areas. Allocating resources in a multi-constellation system with heterogeneous satellite links and highly dynamic user traffic demand poses challenges in ensuring effective resource distribution. To mitigate these complexities and minimize the overhead, there is a growing shift towards utilizing AI for its ability to handle such dynamic problems effectively. This calls for the development of an intelligent decision-making controller using AI to efficiently manage resources in this complex environment.

In the realm of data-driven, real world open datasets play a vital role in the development of AI models addressing radio control optimization problems since the provision of prior information about the several critical parameters of use cases, for instance, how traffic is being changed over a period of time and what are the user terminal demands, could assist in the effectiveness of AI-driven approaches. As a matter of fact, the availability and finding suitable datasets representing the aforesaid properties is cumbersome.

In this context, this deliverable identifies some real-world Open Source (OS) datasets for a subset of the defined use cases in D2.1 and also provides some insights into the realistic traffic pattern and the requirements for the fixed and dynamic user terminals. The aim of gathering and publishing the information of these datasets is to utilize them for further designing the RIC for the use cases under the scope of 6G-NTN. In a nutshell, this deliverable establishes a solid foundation of commercially accessible data, offering the potential to set benchmarks and expedite the optimization of resource allocation problems.

TABLE OF CONTENTS

FIGURES

LIST OF TABLES

Co-funded by the European Union

ABBREVIATIONS

AIS	Automatic Identification System
AI	Artificial Intelligence
CNN	Convolutional Neural Network
CSV	Comma Separated Values
DL	Downlink
DMP	Data Management Plan
DRL	Deep Reinforcement Learning
EC	European Commission
GB	Gigabyte
GIS	Geographic Information System
Kbps	Kilobits Per Second
KPIs	Key Performance Indicators
KM	Kilometer
LAU	Local Administration Units
Mbps	Megabits Per Second
MNO	Mobile Network Operator
Ms	Millisecond
NTN	Non-Terrestrial Networks
OS	Open Source
OSM	Open Street Map
PPO	Proximal Policy Optimization
PPDR	Public Protection and Disaster Relief
QoS	Quality of Service
RIC	Radio Intelligent Controller
RRM	Radio Resource Management
SHP	Shape Files
SNO	Satellite Network Operator
UC	Use Case
UDPP	Urban Data Platform Plus
UE	User Equipment
UTP	User Traffic Profile
UL	Uplink
WGS	World Geodetic System
WP	Work Package

1 INTRODUCTION

This section describes the position of this deliverable in the overall 6G-NTN project activities. Particularly, it delves into a comprehensive discussion in the context of Radio Resource Management (RRM) following the evolution of RRM solutions in NTN aspects.

Subsequently, it illustrates the primary purpose of this deliverable regarding the open datasets within the scope of designing the AI-based RIC in 6G-NTN. This section concludes with the thorough discussion of interconnection of this task with the tasks corresponding to other work packages.

1.1 CONTEXT

D4.3 is explicitly focused on providing the inputs to one of the objectives of 6G-NTN project in terms of OS datasets to design the efficient resource allocation techniques. In the similar context and specifically, in the realm of NTN, the efforts on efficiently utilizing the satellite resources are becoming a priority and requiring attention, both in online and offline scenarios. It is imperative in order to enhance system performance while providing reliable and cost-effective services. The significance of this requirement arises for the multi-orbital and multi-band satellite integrated networks that introduce complexities due to the variability of satellite links and non-uniform user traffic demand as a function of time across diverse use cases, including maritime, PPDR, and direct to handheld, among others [1].

1.1.1 Evolution of Radio Resource Management (RRM) Solutions

To emphasize the importance of resource management problem, significant efforts have been made to optimize the offered capacity of the whole system, taking into account the heterogeneous complexities and by leveraging the recently launched flexible payloads reconfiguration capabilities [2]-[3]. For Instance, in the sense of flexible satellite payload capabilities for the RRM, [4] studied the resource allocation problem and optimized the capacity management of the system using an objective function. In another study, for instance, [5] delves into global resource management by integrating the Quality of Service (QoS) metric and user channel conditions, revealing valuable insights. Meanwhile, [4] introduces a joint optimization tool which simultaneously optimizes the beam width and bandwidth allocation to effectively cater the varying traffic demands.

Nevertheless, the system related constraints and the number of resources to manage at the large scale, these classical methods exhibit limited efficacy in addressing the complexities and thus leading to sub-optimal solutions. Responsively, AI-driven approaches hold significant potential as a key contributor to resource optimization process. In this context, [7] proposes a Deep Reinforcement Learning (DRL) algorithm to optimize the unmet system demand and power consumption using Proximal Policy Optimization (PPO). The authors in [8] mitigates the problem of inter-beam interference in multi-beam satellite system by scheduling the users and allocating bandwidth and power resources intelligently, mainly employing DRL techniques. In another study, the optimal long term capacity allocation plan in a highly complex three-layer heterogeneous satellite network is studied which uses the reinforcement learning as a framework to optimize the capacity allocation but in the realm of synthetic data [9].

1.2 PURPOSE OF THIS DELIVERABLE

It is widely acknowledged that the efficacy of AI paradigm heavily relies on the availability of large and diverse data to effectively learn from the variable environment and provide efficient solutions for resource management. With this understanding, the primary purpose of this deliverable is to identify the real-world datasets with open license, having capabilities to be utilized to design the AI based algorithms that offer practical representations for resource management optimization tasks and can be applied for aforementioned use cases. By offering these datasets, the aim is to gain more practical insights into the realistic and non-uniform behavior of user's resources demand. This, in turn, will enable the development of AI-enabled intelligent decision controller under the scope of 6G-NTN in the practical settings. This AI-enhanced controller can leverage the realistic properties of the datasets and optimize the allocation of resources while serving the maximum users simultaneously.

To gather relevant datasets, a thorough investigation is carried out across a variety of online platforms, including open data portals, online databases, and specialized dataset repositories. Furthermore, the data licensing and documentation are carefully examined to ensure the reliability and credibility of the datasets. This involves thorough examination of the licensing agreements, terms of use, and documentation accompanying the datasets to ensure their suitability for research and analysis purposes. Additionally, it provides the metadata, attributes of each dataset and the guidelines for leveraging open-source information to implement an AI-enabled radio resource management.

1.3 D4.3 RELATION TO OTHER WORKPACKAGES IN 6G-NTN

The general dependencies of all Work Packages (WP) on each other are illustrated in

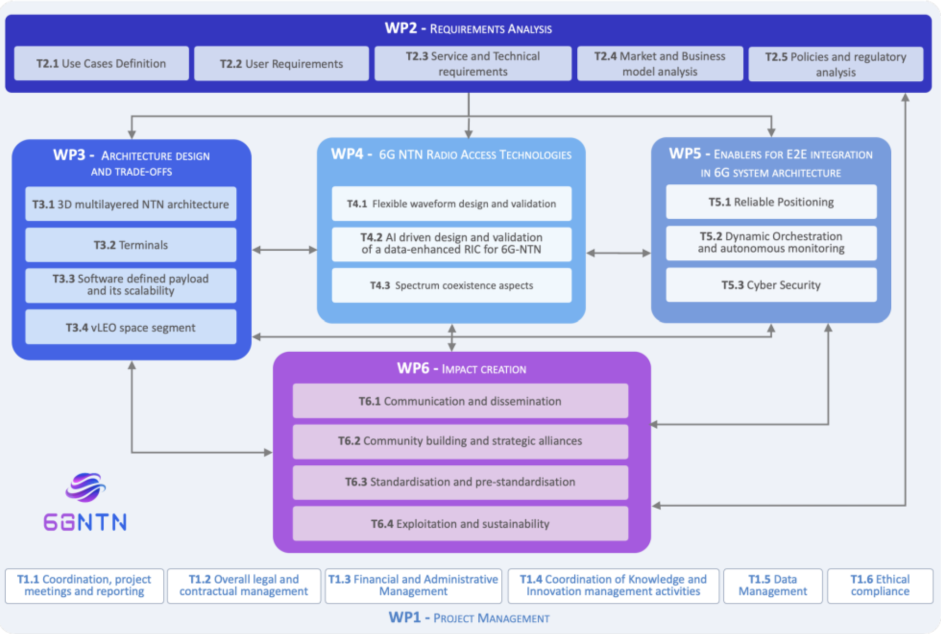


Figure 1. This deliverable is one of the outcomes of WP4 which is related to 6G-NTN radio access technologies which is required to take inputs from the tasks of other WPs and similarly provides outputs to accomplish the other WPs.

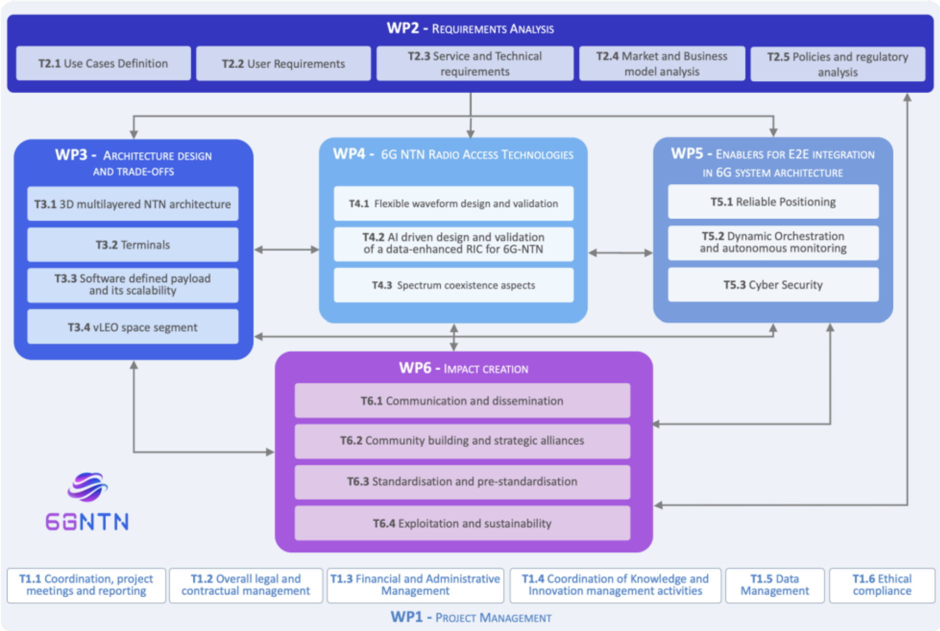


FIGURE 1. 6G-NTN OVERALL WORK PACKAGES ORGANIZATIONS AND THEIR DEPENDENCIES

The organization of task T4.2 with other defined tasks are explicitly given as,

- T4.2 is dependent on the inputs from the task T2.3 of WP2 (Requirements Analysis), related to the provision of service and technical requirements. The T2.3 contributed with the User Traffic Pattern (UTP) values which would be utilized for defining and managing the NTN resources for defined use cases.
- T4.2 might be incorporating 3D multilayered NTN architecture investigated in T3.1 in order to identify the options for deploying the NTN RIC.
- T4.2 considers the outputs of T1.5 for defining the Data Management Plan (DMP).
- T4.2 would contribute to dissemination task T6.1 of WP 6 (Impact Creation) by publishing the work related to RIC in international conferences, workshops, and journals.
- A subset of the outcomes of T4.2 will be developed into a higher TRL in WP5.

The apparent connection of T4.2 with the tasks of other WPs can be seen in Figure 2.

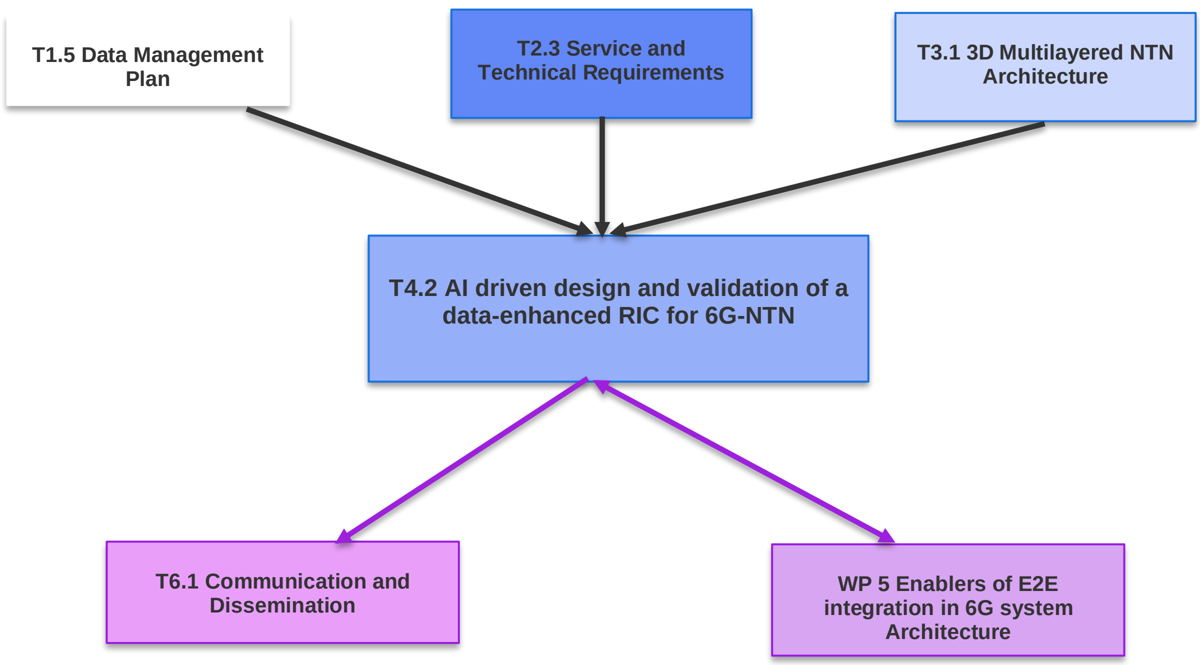


FIGURE 2. THE ASSOCIATION OF TASK T4.2 WITH THE TASKS OF OTHER WPS IN 6G-NTN PROJECT

1.4 ORGANIZATION OF THE DOCUMENT

The document is organized as follows:

Section 2 offers deep insights into the significance of radio resource management and details about the benefits that maritime use case intended to acquire from the 6G-NTN aspects. In this context, it elaborates every aspect of open dataset provided herein such as the creation of the dataset, what attributes it contains, how it can be used for AI- enabled RRM controller and the license information.

Section 3 covers the aspects of PPDR use case, in particular nomadic traffic as one of the examples of PPDR and highlight the significance by leveraging the NTN solutions. Regarding the dataset, it provides the creation description, attributes, guidelines of utilizing the data, and finally the source and license of the available datasets.

Section 4 is about the provision of dataset for direct to smartphone use case and discuss about the critical points of the significance of direct to handheld. Moreover, it provides the guidelines about how one or more datasets can be merged to serve the use case in terms of resource management. Eventually, it offers the very initial insights of performing the beamforming utilizing one of these datasets.

Section 5 draws the conclusion of this deliverable.

2 MARITIME TERMINAL DISTRIBUTION

One of the preeminent datasets for optimizing the resources for the maritime use case, a weeklong dataset with the resolution of one hour, is acquired from a reputable data provider (VesselFinder & VT Explorer) [8]. This dataset consists of realistic marine traffic information updated every one hour obtained from Automatic Identification System (AIS) over the entire Mediterranean Sea. The dataset attributes of maritime traffic are shown in Figure 3.

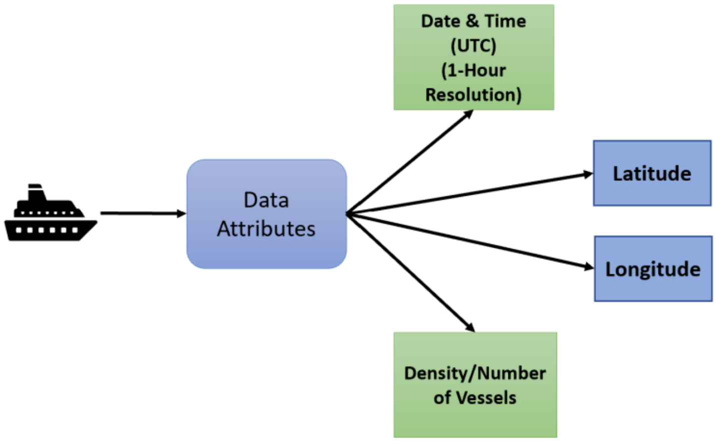


FIGURE 3.THE DATA ATTRIBUTES OF MARINE TRAFFIC DATASET ACQUIRED USING AIS

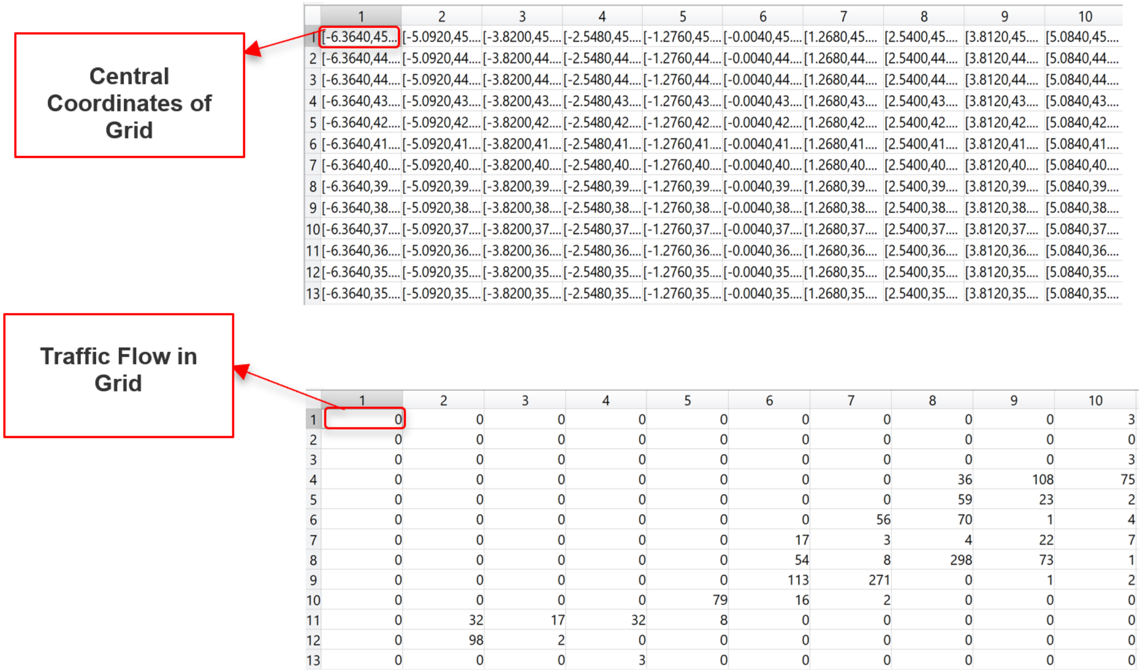
2.1 PURPOSE

It is widely accepted that, concerning the satellite enabled connectivity, NTN has the capability to extend the coverage to vast sea areas, surpassing the limitations of terrestrial networks. In this essence, NTNs hold a significant potential for transforming the maritime industry. This enables seamless connectivity for passenger yachts, facilitates reliable communications among vessels, offshores platforms, maritime surveillance, and ultimately enhances the rescue operations efficiency and allows to effectively meet the safety requirements. Therefore, by leveraging NTN, the maritime industry can significantly benefit from seamless connectivity and communication capabilities.

The purpose of providing the dataset is to be employed for designing the resource management solution to serve the maritime user needs according to the requirements. This facilitates the opportunity to indulge into optimizing the multiple resource allocation strategies; for instance, the available information can be used to train AI models for providing adaptive beamforming and beam allocation, for serving the users based on their throughput demand, flexible user scheduling under the system limitations, and forecasting to better design the system for the next temporal slots etc., which could assist in managing the satellite recourses more effectively.

2.2 FORMAT

The dataset described above is stored in MATLAB files (.mat extension). The entire region of the Mediterranean Sea is converted into grids of haversian distance with 100 km x 100 km resolution. Each hour of the dataset contains two .mat files. One provides a matrix representing the central coordinates (longitude, latitude) of each grid as shown in the upper portion of Figure 4, while the other .mat file give insights into the traffic distribution values within that region as given in the lower portion of Figure 4. Interested users can freely access and explore this dataset for further analysis of testing the different resource allocation solutions or other research purposes without any constraints. To gain a comprehensive understanding of the dataset's structure and to view a sample dataset, refer to the Figure 4 below.



	1	2	3	4	5	6	7	8	9	10
1	[-6.3640,45...	[-5.0920,45....	[-3.8200,45....	[-2.5480,45....	[-1.2760,45....	[-0.0040,45....	[1.2680,45....	[2.5400,45....	[3.8120,45....	[5.0840,45....
2	[-6.3640,44....	[-5.0920,44....	[-3.8200,44....	[-2.5480,44....	[-1.2760,44....	[-0.0040,44....	[1.2680,44....	[2.5400,44....	[3.8120,44....	[5.0840,44....
3	[-6.3640,44....	[-5.0920,44....	[-3.8200,44....	[-2.5480,44....	[-1.2760,44....	[-0.0040,44....	[1.2680,44....	[2.5400,44....	[3.8120,44....	[5.0840,44....
4	[-6.3640,43....	[-5.0920,43....	[-3.8200,43....	[-2.5480,43....	[-1.2760,43....	[-0.0040,43....	[1.2680,43....	[2.5400,43....	[3.8120,43....	[5.0840,43....
5	[-6.3640,42....	[-5.0920,42....	[-3.8200,42....	[-2.5480,42....	[-1.2760,42....	[-0.0040,42....	[1.2680,42....	[2.5400,42....	[3.8120,42....	[5.0840,42....
6	[-6.3640,41....	[-5.0920,41....	[-3.8200,41....	[-2.5480,41....	[-1.2760,41....	[-0.0040,41....	[1.2680,41....	[2.5400,41....	[3.8120,41....	[5.0840,41....
7	[-6.3640,40....	[-5.0920,40....	[-3.8200,40....	[-2.5480,40....	[-1.2760,40....	[-0.0040,40....	[1.2680,40....	[2.5400,40....	[3.8120,40....	[5.0840,40....
8	[-6.3640,39....	[-5.0920,39....	[-3.8200,39....	[-2.5480,39....	[-1.2760,39....	[-0.0040,39....	[1.2680,39....	[2.5400,39....	[3.8120,39....	[5.0840,39....
9	[-6.3640,38....	[-5.0920,38....	[-3.8200,38....	[-2.5480,38....	[-1.2760,38....	[-0.0040,38....	[1.2680,38....	[2.5400,38....	[3.8120,38....	[5.0840,38....
10	[-6.3640,37....	[-5.0920,37....	[-3.8200,37....	[-2.5480,37....	[-1.2760,37....	[-0.0040,37....	[1.2680,37....	[2.5400,37....	[3.8120,37....	[5.0840,37....
11	[-6.3640,36....	[-5.0920,36....	[-3.8200,36....	[-2.5480,36....	[-1.2760,36....	[-0.0040,36....	[1.2680,36....	[2.5400,36....	[3.8120,36....	[5.0840,36....
12	[-6.3640,35....	[-5.0920,35....	[-3.8200,35....	[-2.5480,35....	[-1.2760,35....	[-0.0040,35....	[1.2680,35....	[2.5400,35....	[3.8120,35....	[5.0840,35....
13	[-6.3640,35....	[-5.0920,35....	[-3.8200,35....	[-2.5480,35....	[-1.2760,35....	[-0.0040,35....	[1.2680,35....	[2.5400,35....	[3.8120,35....	[5.0840,35....

	1	2	3	4	5	6	7	8	9	10
1	0	0	0	0	0	0	0	0	0	3
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	3
4	0	0	0	0	0	0	0	36	108	75
5	0	0	0	0	0	0	0	59	23	2
6	0	0	0	0	0	0	56	70	1	4
7	0	0	0	0	0	17	3	4	22	7
8	0	0	0	0	0	54	8	298	73	1
9	0	0	0	0	0	113	271	0	1	2
10	0	0	0	0	79	16	2	0	0	0
11	0	32	17	32	8	0	0	0	0	0
12	0	98	2	0	0	0	0	0	0	0
13	0	0	0	3	0	0	0	0	0	0

FIGURE 4. THE AVAILABLE DATASET STRUCTURE OF MARITIME TRAFFIC FLOW. THE UPPER FIGURE CORRESPONDS TO CENTRAL COORDINATES OF EACH GRID WHILE THE LOWER FIGURE IS RELATED TO TRAFIC FLOW IN EACH GRID AT SPECIFIC HOUR.

2.3 SOURCES

The user positions for the marine data are acquired by VesselFinder & VT Explorer [8]. Since the position data is captured by various land-based AIS stations and satellites, therefore, it becomes one of the accurate and reliable sources of acquiring the dataset.

2.4 CREATION DESCRIPTION

The maritime density map dataset is created by processing, the raw data acquired from the vessel finder in comma separated values (.CSV) format in MATLAB. The Figure 5 illustrates a snapshot of the dataset represented as a density grid map. The grid map is a function of longitude and latitude where each grid encompassing an area of 100 km x 100 km using haversine distance. The haversian distance is chosen due to its capability of taking the curvature of Earth´s surface into account, resulting in more accurate calculation as compared to simpler methods such as Euclidean distance. The haversine distance can be calculated using the given equations.

$$a = sin^2\left(\frac{\Delta\emptyset}{2}\right) + cos\emptyset_1 . cos\emptyset_2 . sin^2\left(\frac{\Delta\lambda}{2}\right) \quad (1)$$

$$c = 2. atan2(\sqrt{a}, \sqrt{1-a}) \quad (2)$$

$$d = R.c \tag{3}$$

where R is the earth radius, i.e., 6371 km, $\Delta\emptyset$ and $\Delta\lambda$ are the absolute difference between the latitudes and longitudes, respectively, and d is the haversine distance. The snapshot density grid map as a function of geocoordinates is shown in Figure 5.

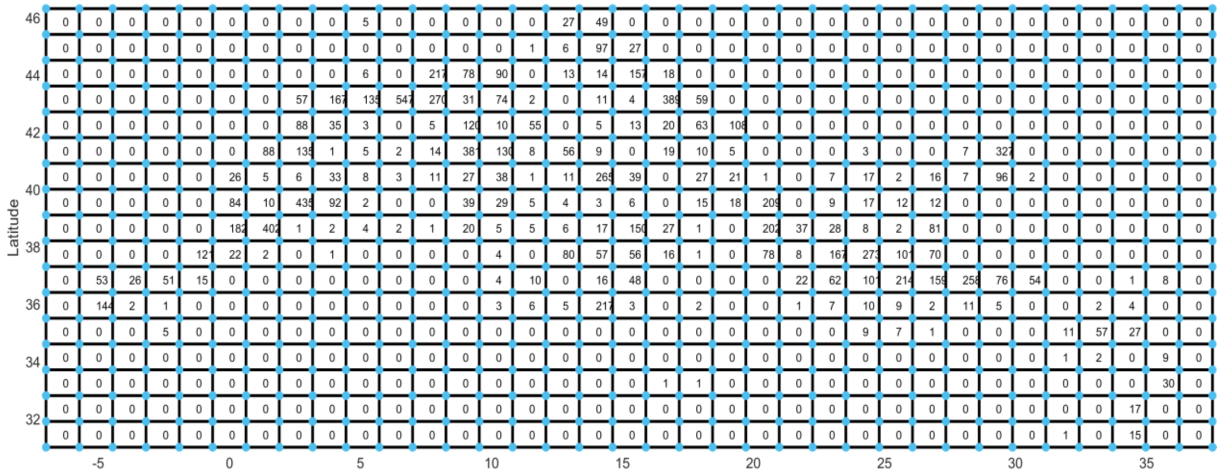


FIGURE 5. THE SNAPSHOT OF MARINE TRAFFIC DENSITY OF AN HOUR DURATION

To construct this grip map, the latitudes of Mediterranean Sea are evenly spaced with a factor of 0.9, which is equivalent to the haversine distance of 100 km with respect to latitudes while keeping the longitude constant. This process determines the vertical lines of a single gird. Subsequently, the longitudes of the Mediterranean Sea are evenly spaced with a factor of 1.2, which is equivalent to the haversine distance of 100 km with respect to longitudes, while maintaining the latitudes constant, resulting in horizontal lines of the grid. By combining these vertical and horizontal lines, a complete grid map is formed.

The numerical value within each grid cell represents the number of vessels observed within the corresponding 100 km x 100 km area. This grid-based representation allows for a comprehensive analysis and visualization of maritime traffic variations within the dataset, offering insights into spatio-temporal distribution across different geographical regions of the Mediterranean Sea.

Remarkably, the original dataset contains information regarding the vessel type (e.g. cargo, passenger,…). In this context, depending on the radio management functions developed in T4.2, this dataset can be updated with a lower spatial granularity and more information regarding the vessel type in order to infer traffic demand per vessel.

2.5 LICENSE

CTTC License: The maritime traffic flow dataset (processed files) is accompanied by permissive license that allow users to access, commercially use, modify, and distribute the data freely. However, it is essential to give proper credits to the creators of the dataset (CTTC) as a sign of acknowledgement by referencing the 6G-NTN and this deliverable.

2.6 METADATA

The metadata of the maritime density grid map provides essential information about the data attributes, data types, and the description of each attribute, including geographical coordinates, traffic flow, and time stamps. Such details enable the users to process and handle the datasets for data analysis in a more convenient. The metadata is shown in Table 1

No.	**Attributes**	**Description**	**Data Type**
1	Latitude	The latitude of the center point of grid map	Double
2	Longitude	The longitude of the center point of a grid map	Double
3	Density/Traffic Flow	The number of vessels in a particular grid	Unsigned integer
4	Time	The time information at which the snapshot taken [hh:mm]	String

TABLE 1.THE METADATA OF MARITIME DENSITY MAP

3 PPDR OR TEMPORARY EVENTS TERMINAL

Obtaining the dataset for the Public Protection and Disaster Relief (PPDR) use case can be challenging due to the difficulties in identifying the areas with limited terrestrial communication infrastructure where the network performance is not up to the mark to serve even for temporary events. Moreover, personal data privacy and security laws impose strict obligations on the collection, storage and publication of any data related to people in the EU, including their mobility and geolocalization, in complying with the General Data Protection Regulation (GDPR) or other national regulations (e.g., issued by the CNIL in France).

In this context, there are two potential approaches for acquiring the datasets for 6G-NTN. The first option involves conducting surveys and connectivity tests to generate the dataset from scratch. However, this approach can be resource-intensive and time-consuming. Alternatively, utilizing an existing dataset generated by the other entities through the similar way is another possibility but it is often challenging to find datasets with open licenses. In a general manner, there is no open dataset for realistic “temporary events or PPDR terminal distribution” equivalent to the ones provided by VesselFinder or VT Explorer, for realistic marine traffic information and neither anything is openly available that can be further process to forecast the distribution.

The user distribution could be potentially derived from public datasets from national census [9],or statistics on local & regional population densities. However, such information is static and does not reflect the utilization of connectivity services, nor the effective user mobility (residential areas are often quite empty during week days, as well as industrial and commercial zones at night or Sundays). As of today, nearly everybody is equipped with smartphones and, Mobile Network Operators (MNOs) are often perceived as a potential provider of population density live data and some have developed new services [10] allowing, for example, the optimization of road engineering or truck traffic management, while being fully compliant with GDPR, in particular thanks to anonymization. However, such services remain commercial and related data cannot be made public. Finally, MNOs and Satellite Network Operators (SNOs) manage other types of datasets, representing the traffic evolution at a fine granularity (e.g., data per base station, every 15minutes). However, such data being very sensitive and strategic, it could not be made available for 6G-NTN.

Fortunately, Ookla´s open data initiative facilitates with a trusted repository that offers accessible and modifiable data that enables to adopt the data based on the objectives [11]. Furthermore, the Urban Data Platform Plus (UDPP), provided by European Commission (EC) put additional efforts to Ookla´s Speedtest data, by offering the valuable insights into broadband speed for both fixed and mobile networks at the municipalities level across the European region [12] making it easier to work.

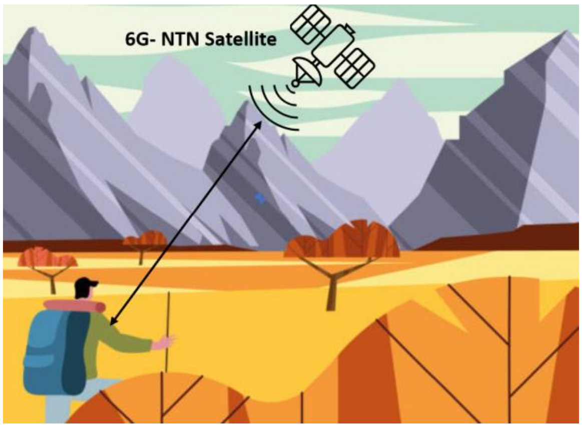
3.1 PURPOSE

Taking into account the general term of PPDR or temporary event, nomadic traffic can be one of the examples of PPDR event that involves the movement of travelers and vehicles across different locations and in need of internet connectivity in remote or underserved environments such as mountainous and isolated islands. Additionally, addressing the communication needs of nomadic traffic is not commonly prioritized within existing communication system. Therefore, NTN emerges as a valuable support in the realm of reliable, ubiquitous connectivity and seamless communication for nomadic travelers on the pause and on the move. The utilization of 6G-NTN enables the travellers to access real time navigation system, location tracking and internet connectivity even in remote or challenging environment. This become particularly crucial in emergency situations where a reliable communication is essential. Taking into

Co-funded by the European Union

account these factors, nomadic traffic can be one of the potential examples of PPDR to be supported by non-terrestrial communication systems since these systems have the potential to play a vital role to enhance safety, efficiency and convenience for the travellers on the move.

The main idea is to identify the data that provides insights into the underserved areas and areas where the terrestrial infrastructure is only partially accessible or not available at all by looking into network performance that is below a specific threshold. This throughput information can be extracted from the aforementioned datasets which provide insights about the resources required through NTN. Additionally, these datasets could be used as an input to AI- enabled resource management controller to manage the resources only for the identified geographical areas where the seamless connectivity is required. The representation of a nomadic user can be seen in Figure 6.



.

FIGURE 6.THE NOMADIC USER CONNECTING WITH 6G-NTN SATELLITE IN UNDERSERVED AREA.

3.2 FORMAT

In case of Ookla´s data, the dataset is provided into two formats: shape (SHP) files and Apache Parquet format. Shape files are easily be supported and processed in Geographic Information System (GIS) software while the Parquet files are suitable for processing using MATLAB and Python programming languages [11].

Regarding the other dataset from EC UDPP, the dataset follows the .CSV format [12], which is more user friendly to process information at municipality level and adaptable for designing the resource management controller since it is already processed in a way that can be required to be an as input to train AI models.

3.3 SOURCES

The position coordinates of the areas and the corresponding network performances are explicitly provided by Ookla speed test platform for both fixed and mobile networks globally.

3.4 CREATION DESCRIPTION

This Ookla´s open dataset facilitates with the global fixed broadband and mobile network performance in the form of tiles. The fixed broadband refers to measurements taken from mobile devices with non-cellular type connection i.e., Wi-Fi or Ethernet, whereas the mobile network measurements are taken from mobile devices with cellular type connection, for instance, 4G LTE and 5G NR. The tiles are partitioned into zoom level 16 Web Mercator projection, measuring approximately 610.8 meters by 610.8 meters at the equator. The zoom level has a meaning of covering the earth region, for instance, zoom level 0 gives the size of tile equivalent to entire globe while zoom level 1 represents into 4 tiles. The dataset is provided into two formats: shape (SHP) files and Parquet files with each tile represented in world Geodetic System (WGS- EPSG:4326). The dataset spans from the period of Q1 2019 to the recently complete quarter Q2 2023 and being updated every quarter year (three months) [11]. Hence, a continuous updated information can be acquired after a defined interval of time. The tests are performed at random hours with several devices, or a single device performs multiple tests in some cases. The measurements given by all devices and tests are then aggregated to get the average value in each tile. Therefore, the provided key performance indicator (KPIs) such as average latency and speed for both upload and download are based on devices participation and tests repetitions which is related to the number of devices is used to perform the tests of acquiring the statistics. Therefore, higher the number of devices and tests lead to more accurate values of KPIs.

On the other hand, EC UDPP data is a further processing of Ookla dataset to convert into the municipality level scale that identifies the average download speed in each municipality rather than the zoom level tiles where the municipality boundaries are defined by local administration units (LAU) [12].

However, the relevance of this data to the 6G-NTN for nomadic traffic lies in its sparse nature since many geographical regions within the dataset exhibit no statistics at all. Hence, the idea of adopting the Ookla´s Open dataset for is to exploit the property of being sparse and on top of that, there are many tiles around the globe where the network performance is inadequate. Therefore, for the regions with the poor network performance and no coverage, the implementation of 6G-NTN becomes necessary to fill the gaps and improve connectivity along with the management of resources based on requirements. The numerical data provided by these datasets possesses sufficient capability to be utilized as input for training AI- enabled learning networks which can optimize the RRM controller for the provision of connectivity in accordance with the specific requirements of the geographical regions.

3.5 LICENSE

Ookla Dataset License: The Ookla dataset is licensed by Attribution- Non-Commercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) that gives the by permissive license that allow users to access, modify and redistribute the data freely in any format. However, it couldn´t be used for commercial purposes [13].

EC UDPP License: The EC Urban Data Platform Plus is licensed by Attribution 4.0 International (CC BY 4.0) that gives the by permissive license that allow users to access, modify and redistribute the data freely in any format, provided appropriate credit is given and changed are indicated [14].

3.6 METADATA

The metadata of global fixed and mobile network performance data map provided by Ookla give insights into the essential information about the network performances for both fixed and mobile through various KPIs. The metadata in Table 2.

No.	Key Performance Indicators (KPIs)	Description	Data Type

1	**Tile**	The coordinates of the tile	Numeric
2	**avg_d_kbps**	The average download speed in kbps*	Integer
3	**avg_u_kbps**	The average upload speed in kilobits per second	Integer
4	**avg_lat_ms**	The average latency measured in ms*	Integer
5	**avg_lat_d_ms**	The average latency during the download phase measured in ms	Integer
6	**avg_lat_up_ms**	The average latency during the upload phase measured in ms	Integer
7	**Tests**	The number of tests taken in the tile.	Integer
8	**Devices**	The number of unique devices contributing tests in the tile.	Integer
9	**Quadkey**	The quadkey representing the tile.	String

*kbps= kilobytes per second, ms= millisecond

TABLE 2.THE METADATA OF GLOBAL FIXED AND MOBILE NETWORK PERFORMANCE DATAMAP BY OOKLA [11].

The metadata of the data map offered by EC UDPP is originally a processed version of Ookla dataset but summed up at municipality level it gives more accurate insights into the regional level information about the network performances for both fixed and mobile through the KPI such as average download speed corresponding to municipality geographical coordinates. The municipality boundaries are those defined by LAU.

The metadata table for EC UDPP dataset is in Table 3.

No.	Attributes	Description	Data Type
1	**Territory ID**	The identifier of the Territory	String
2	**Municipality Name**	The municipality name with Geographical Coordinates.	String
3	**Average speed**	The average download speed in Mbps*	Double
4	**Year**	2018-2022	Integer

Mpbs = Megabytes per second

TABLE 3.THE METADATA TABLE FOR EC URBAN DATA PLATFORM PLUS DATASET [12]

4 DIRECT-TO-SMARTPHONE TERMINAL AND TRAFFIC DISTRIBUTION

Based on the identification and the specified use case criteria, the task of finding the suitable and readily available dataset for direct-to-smartphone is deemed extremely challenging. However, one approach could be to merge the multiple datasets to extract a more accurate representation of the requirements for direct-to-smartphone use case considering both temporary (disaster situations) and permanent (remote and unserved areas) events. In this essence, the strategic integration of the Ookla or EC UDPP dataset, as detailed in preceding subsection alongside the population density dataset provided by Meta [15], presents an opportunity to derive insights about the requirements of NTNs system design to better serve the direct-to-handheld use case.

4.1 PURPOSE

A direct-to-smartphone use case within the realm of 6G NTN is of great significance, particularly in the situations when the connectivity is needed through satellite networks for various temporary and permanent purposes. For the permanent purpose, the need for NTN becomes crucial for this use case when considering the provision of connections in remote villages that lack the overall terrestrial infrastructure. In such scenario, NTN enables users to access the internet, make voice calls and exchange text messages. Another prominent but temporary scenario is during emergencies, for instance, natural disasters where the terrestrial network may be severely impaired. This enables the relief for the affected individuals to seek emergency assistance and communicate about their situation to authorities. Similarly, it empowers the authorities in the affected areas to communicate and expedite the rescue operations in the absence of terrestrial facilities.

In light of the above discussion, the objective of this dataset is to offer information regarding potential positions of direct-to-smartphone users and a insights of the traffic demands of those users.

4.2 FORMAT

The format of Ookla dataset is already discussed in the preceding section while the population density dataset follows the .CSV format that gives the (statistical) number of people at the given geographical coordinates [15].

4.3 SOURCES

The statistical number of people and the corresponding geographical positions are explicitly provided by Meta database whereas the coordinates of regions that are potential in need of NTN service are provided by Speed Test Platform as discussed in previous section.

4.3.1 User Traffic demand

The information within the datasets provided for direct-to-smartphone use case aims to support the design of radio network management functions. However, to accurately manage the resources, it is essential to have an important component i.e. per user terminal demand, for these use cases. Although, the two datasets can help to identify the underserved areas and user terminals density, they lack details about the per user terminal demand in those remote and underserved areas where terrestrial networks might be accessible with limited capacity or no accessibility at all.

In this context and regarding the use cases, in particularly PPDR and direct-to-smartphone, this document provides the realistic traffic pattern for the purpose of integration with the provided datasets to complete the input information typically required for designing the AI-enabled controller. The data, which refers to user demand as User Traffic Profile (UTP), is a collection of KPIs, whose values provide information about the desired emergent features of the other provided datasets. As an example, Table 4 outlines this set of KPIs for the direct-to-smartphone or PPDR use case that comes directly from D2.3 '*Report on system requirements*'. In this regard, a UTP encompasses KPIs such as latency, information on link availability and reliability, experienced user data rate and activity factor [16], among other performance targets. Their values are derived from the analysis of several direct-to-smartphone subcases. For instance, the UTP outlined in Table 4 relates to a use case whereby users are trekking in areas without continuous terrestrial coverage and wish to continue using their audio chat and video streaming services, both in outdoor and light indoor conditions, with different quality of experience. This specific subcase notably highlights high user uplink (UL) requirements, which can be seen via the high values on the uplink Experienced User Data Rate and Activity Factor. It completes a more general case in which the Activity Factor downlink (DL) is significantly higher than the UL.

Direct-to-Smartphone	UE type	Experienced user Data Rate		Activity Factor		Monthly User Traffic		Busy Hour Usage Rate		Latency	Radio Link Availability	Reliability	Mean User Density	Area Traffic Capacity		User Distribution	UE Speed
		DL	UL	DL	UL	DL	UL	DL	UL					DL	UL		
	(Text)	Mbps		%		GB/Month		Kbps		ms	%	%	user/km^2	Kbps/km^2		(Text)	km/h
Outdoor	Handheld	2	10	5.0	5.0	167	33	100	500	< 30	99.9	99.999	5	500	2500	UEs distributed on trails	20
Light indoor	Handheld	0.01	0.01	5.0	5.0			0.5	0.5	< 50	99.9	99.99	5	2.5	2.5		

TABLE 4.THE USER TRAFFIC PROFILE FOR DIRECT-TO-SMARTPHONE USE CASE

The description of KPIs is detailed as follows [16],

- **Experienced User Data Rate**: Represents the minimum data rate needed to achieve a target Quality of Experience (QoE).
- **Activity Factor**: Ratio of simultaneous active UEs to the total number of UEs.
- **Monthly User Traffic:** Monthly User Traffic is expressed in GB/user/month for both the downlink and uplink.
- **Busy Hour Usage Rate**: Represents as,

  Busy Hour Usage Rate DL = Experienced Data Rate DL * Activity Factor DL

  Busy Hour Usage Rate UL = Experienced Data Rate UL * Activity Factor UL
- **Mean User Density:** Represents the average user density over a defined area.
- **Area Traffic Capacity:** Total traffic throughput served per geographic area.

Disclaimer: These UTP values are the inputs from D2.3, which is being updated iteratively. Therefore, these values are for illustration and subject to change. Interested users can refer to finalized values in D2.3 Section 4.3 (Performance Requirements using User Traffic Profiles).

Co-funded by the European Union

4.3.2 Creation description

The acquisition of high-resolution population density map involves employing state of the art computer vision techniques on high resolution (50 cm per pixel) satellite imagery data integrated with the publicly available census data. The idea behind building the map for population density involves employing the Convolutional Neural Network (CNN) to identify the buildings from public accessible mapping service such as OpenStreetMap (OSM). Subsequently, around 100 million labeled examples are extracted and added into the training set to train the ResNet50 model. The population density map exhibits impressive accuracy level, providing data with 30 x 30-meter grid resolution, making it reliable and authentic dataset for utilization [15].

Since, the meta dataset offers the statistical information regarding the population count at each geocoordinate while on the other hand, the Ookla dataset presents data on fixed and mobile network performance for the same geocoordinates. By leveraging the population density from the Meta´s dataset and then examining the mobile network performance data from Ookla dataset, the valuable insights can be obtained regarding the required resources. Additionally, per terminal demand is provided for the direct to smartphone use case which is also valid for the nomadic use case. Integrating all this information based on geographical coordinates from these sources allow to estimate the satellite resources such as bandwidth required, throughput capacity etc., that NTN should allocate to meet the unserved demand based on population. The idea to integrate is shown in Figure 7.

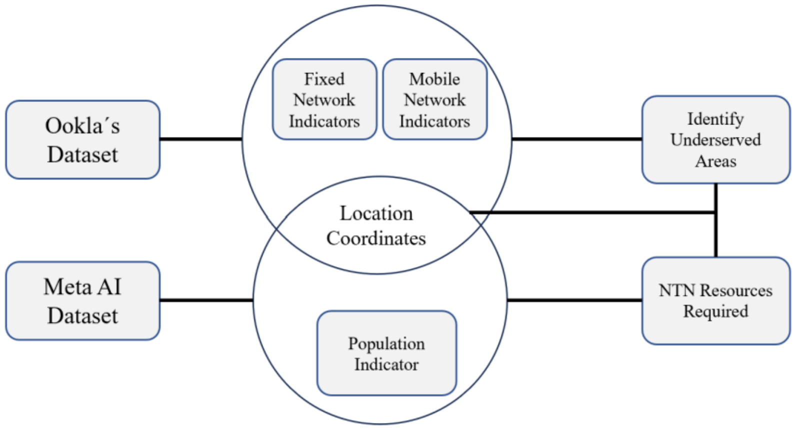


FIGURE 7.THE INTEGRATION STRATEGY FROM OOKLA AND META AI POPULATION DENSITY DATASETS

4.4 LICENSE

The processed datasets have the same license of the ones used for creating them. In case colliding license terms arise, we opt to consider an exclusive license to be used in the context of the project.

4.5 METADATA

The metadata of population density data map provided by Meta AI give insights into the essential information about the statistical number of people at each coordinate with the resolution of 30 x 30 m grid, enabling users to process and handle the datasets accordingly. The metadata are listed in Table 5.

No.	Attributes	Description	Data Type
1	Latitude	The latitude information of 30 x30 meter grid	Double
2	Longitude	The longitude information of 30 x30 meter grid	Double
3	Population Density	The (statistical) number of people in the grid	Double

TABLE 5.THE METADATA OF POPULATION DENSITY DATAMAP BY META AI [15]

4.6 BEAMFORMING EXAMPLE USING INTEGRATED DATASET

Considering the integration of Ookla and Meta AI datasets for direct-to-smartphone use case, a beamforming example of utilizing the integrated dataset is provided. This example offers initial insights into resource management through NTN system based on user terminal requirements outlined in Table 4. The example showcases the projection of a beam through satellite to serve the users on the region where NTN accessibility is truly required. The idea involves initially identifying the regions through Ookla speed test dataset where the average speed falls below 2 Mbps (Megabits per second). The minimum data rate is an essential requirement by user terminal for outdoor scenario as indicated in Table 4. Additionally, the similar minimum requirements can be considered when terrestrial network accessibility is crumbled under the emergency conditions.

In this essence, the 119 tiles from the Ookla dataset are identified within the vicinity of Barcelona where the average speed does not meet the requirements. The geocoordinates corresponding to these tiles are then extracted for further processing. Subsequently, the statistical population density is accumulated for the selected geocoordinates as chosen by using Speed test dataset.

Once the total area for those 119 tiles is identified, the beam is projected at the mean point (41.6289 Latitude, 2.2568 Longitude) of the identified region from the simulated satellite. Based on the user terminal pattern experienced in Table 4 and the number of statistical amount of people, the overall demand at a given time can be estimated roughly and hence, beamforming technique can be applied to project a beam with specific parameters to fulfil the requested throughput. The beam projected on the identified vicinity of Barcelona is depicted in Figure 8.

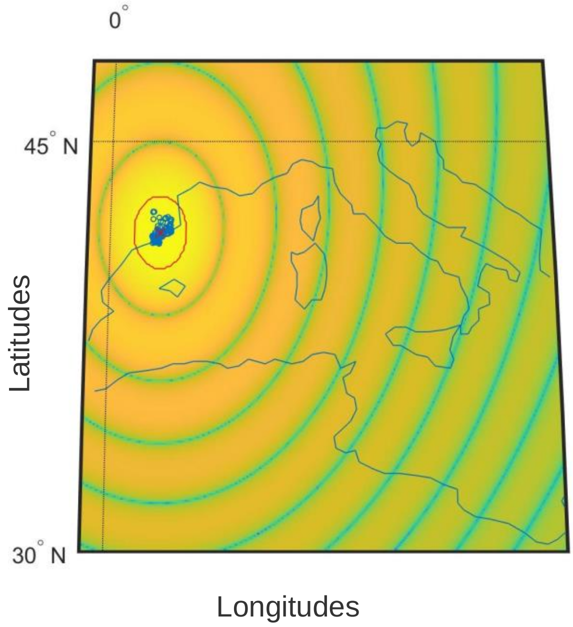


FIGURE 8. THE BEAM PROJECTED ON THE IDENTIFIED VICINITY OF BARCELONA

The red circle is the beam projected by the satellite and blue points are center points of those tiles where the average speed is below 2 Mbps. The accumulated user demand based on the outcomes acquired from population density dataset at the same coordinates as in Speed test dataset is illustrated in Table 6.

No.	**Population intended to serve (Percentage)**	**Accumulated User Demand at Identified Region (Gbps)***
1	10	1018
2	7	712
3	5	509
4	3	305

TABLE 6. ACCUMULATED USER DEMAND BASED ON POPULATION

**The population need to be served simultaneously.

The example presented here is on smaller scale, however, for a large scale when there are multiple areas are identified as potential regions for NTN services, an AI based adaptive beamforming technique could be applied. This technique will project the beams based on clustering on the areas according to the identified tiles. Since the user requirements will already be known, the system can be designed to provide the sufficient throughput to serve all the users within the identified regions.

5 CONCLUSIONS

The management of radio resources in NTN is of utmost importance as it directly affects the overall capacity of the communication system. In this context, efficiently allocating resources in a multi-orbital and multi-satellite system becomes challenging due to the variability of inter and intra satellite links as well as the heterogeneous user traffic demand. Coping with these challenges requires to incorporate these variabilities in designing the resource management techniques. To do so, there is a growing shift towards leveraging AI for its effective abilities to handle such problems effectively in comparison with classical algorithms. Since AI algorithms heavily rely on datasets, this document mainly focused on the identification and the provision of open datasets as potential candidates for AI-enabled radio resource controller under the scope of 6G-NTN project. Such a controller can effectively manage the satellite resources for multiple use case such as Maritime Coverage for search and rescue coast guard intervention, Adaptation to PPDR or Temporary Events, and Consumer Handheld Connectivity and Positioning in Remote Areas.

In the domain of open datasets, this document provides access to the identified datasets and offers a comprehensive overview of their relevance for each use case. It provides valuable insights into how these datasets can be effectively utilized in managing the radio resources by taking into the account the given real-world values of the traffic flow, user traffic demand. Furthermore, utilizing the datasets information, this document helps in identifying the remote and underserved areas where terrestrial infrastructure is not accessible and therefore, 6G-NTN emerges as a prime candidate to be designed for such scenarios, providing reliable communication and seamless connectivity.

The preceding sections showed a robust source of datasets able to address a large subset of 6G-NTN use cases. Forthcoming 6G-NTN terminal positions are properly inferred from real data and further datasets could be obtained based on the spatial distribution. Despite no real traffic demand traces are provided, educated guesses from real situations are provided.

REFERENCES

[1] Al-Hraishawi H, Lagunas E, Chatzinotas S. Traffic simulator for multibeam satellite communication systems, 10th Advanced Satellite Multimedia Systems Conference and the 16th Signal Processing for Space Communications Workshop (ASMS/SPSC) 2020, pp. 1-8.

[2] Thale Space Inspire satellites, (https://www.thalesgroup.com/en/markets/space/space-connect)

[3] Airbus OneSat satellites, (https://www.airbus.com/en/space/telecom/onesat)

[4] G. Cocco, T. de Cola, M. Angelone, Z. Katona and S. Erl, "Radio Resource Management Optimization of Flexible Satellite Payloads for DVB-S2 Systems," in IEEE Transactions on Broadcasting, vol. 64, no. 2, pp. 266-280, June 2018.

[5] Su Y, Liu Y, Zhou Y, Yuan J, Cao H, Shi J. Broadband LEO satellite communications: Architectures and key technologies. IEEE Wireless Communications (2019), 26(2), pp 55-61.

[6] Honnaiah, P.J., Maturo, N., Chatzinotas, S., Kisseleff, S. and Krause, J., 2021. Demand-based adaptive multi-beam pattern and footprint planning for high throughput GEO satellite systems. IEEE Open Journal of the Communications Society, 2, pp.1526-1540.

[7] J. Luis, M. Guerster, et al., “Deep reinforcement learning for continuous power allocation in flexible high throughput satellites,” 2019 IEEE Cognitive Communications for Aerospace Applications Workshop (CCAAW), 2019, pp. 1– 4.

[8] T. Leng, Y. Wang, D. Hu, G. Cui and W. Wang, "User-level scheduling and resource allocation for multi-beam satellite systems with full frequency reuse," in China Communications, vol. 19, no. 6, pp. 179-192.

[9] Jiang, C. and Zhu, X., 2020. Reinforcement learning based capacity management in multi-layer satellite networks. IEEE Transactions on Wireless Communications, 19(7), pp.4685-4699.

[10] The FREE AIS vessel tracking platform (https://www.vesselfinder.com/).

[11] Public Datasets from National Cencus (https://www.insee.fr/en/information/2517226)

[12] Flux Vision: The mobile data analytics solution, statistic indicator of attendance, pattern of movement (ttps://www.orange-business.com/en/solutions/data-intelligence-iot/flux-vision).

[11] Speedtest by Ookla Global Fixed and Mobile Network Performance Maps was accessed from (https://registry.opendata.aws/speedtest-global-performance).

[13] The European Commission Urban Data Platform Plus (https://urban.jrc.ec.europa.eu/).

[14] The license of Ookla Global Speed Test https://registry.opendata.aws/speedtest-global-performance/.

[15] The license of EC Urban Platform Plus Database https://commission.europa.eu/legal-notice_en.

[16] The country wise population density dataset using AI, provided by META https://dataforgood.facebook.com/dfg/docs/methodology-high-resolution-population-density-maps.

[17] "Service requirements for the 5G system," 3GPP Technical Specification TS22.261, stage 1 (Rel. 19), v19.3.0, June 2023.

Co-funded by the European Union