Horizon 2020

OLGA

D2.3: IT SOLUTION FOR MULTIMODALITY AND TRAFFIC FLOW OPTIMIZATION, DESIGN AND VALIDATION

Grant Agreement number: 101036871

Project acronym: OLGA

Project title: HOListic & Green Airports

Funding scheme: Innovation Action (IA)

Start date of the project: 1st October 2021

Duration: 60 months

Project coordinator: Virginie Pasquier, Project Manager Environment / Energy, ADP

Tel: +33 7 88 35 16 07

E-mail: virginie.pasquier@adp.fr

Project website address: www.olga-project.eu

OLGA project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement n° 101036871

DOCUMENT INFORMATION

Document Name	**IT solution for multimodality and traffic flow optimization, design and validation**
Version	V1
Version date	14/02/2025
Authors	Krešimir Vidović (ENT), Saša Vojvodić (ENT), Robert Radović (ENT), Elizabeta Grahovac (ENT), Miroslav Zver (ENT), Ana Najev (ENT), Domagoj Leljak (ENT)
Security	Public

APPROVALS

	Name	**Company**	**Date**	**Visa**
Coordinator	Virginie Pasquier	ADP	03/02/2025	OK
WP Leader	Marin Tica	MZLZ	07/10/2024	OK
Task Leader	Saša Vojvodić	ENT	03/10/2024	OK
Quality Manager	Melike Riollet	LUP	14/02/2025	OK

DOCUMENT HISTORY

Version	**Date**	**Modification**	**Authors**
V01	01/08/2024	Initial version	Krešimir Vidović (ENT), Saša Vojvodić (ENT), Elizabeta

			Grahovac (ENT), Miroslav Zver (ENT), Ana Najev (ENT)
V02	26/09/2024	Integrated, update version	Krešimir Vidović (ENT), Saša Vojvodić (ENT), Elizabeta Grahovac (ENT), Miroslav Zver (ENT), Ana Najev (ENT), Ana Najev (ENT), Domagoj Leljak (ENT)
V03	30/09/2024	Abbreviation list added	Krešimir Vidović (ENT), Saša Vojvodić (ENT)
V04	03/10/2024	Technology Readiness Level assessment chapter added	Krešimir Vidović (ENT), Saša Vojvodić (ENT)
V1	14/02/2025	Quality review	Melike Riollet (LUP)

Table of Contents

List of tables

List of figures

List of Abbreviations

A/D	Arrival/Departure
AI	Artificial Intelligence
API	Application Programming Interface
B2B	Business to Business
CDG	Charles De Gaulle airport
CDR	Call data record
CO2	Carbon dioxide
CO2e	Carbon dioxide equivalent
CRM	Customer relation management
CSV	Comma-separated values
DATEX	European standard for traffic and travel information
ETS	Emissions Trading System
EU	European Union
E-UTRA	Evolved UMTS Terrestrial Radio Access
FRAME	European Intelligent Transport Systems (ITS) Framework Architecture

GDPR	General Data Protection Regulation
GHG	Greenhouse gases
GIS	Geographic Information System
GP	Main Parking
GPS	Global Positioning System
GTFS	General Transit Feed Specification
HDFS	Hadoop Distributed File System
HTTP	Hypertext Transfer Protocol
IMSI	International Mobile Subscriber Identity
ISO	International Organization for Standardization
IT	Information technology
ITS	Intelligent Transport Systems
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
LTE	Long-Term Evolution; a fourth-generation (4G) wireless standard
MCC	Mobile Country Code
ML	Machine Learning
MQTT	Message Queuing Telemetry Transport
MZLZ	Međunarodna zračna luka Zagreb, Zagreb Airport
NETEX	Network Timetable Exchange
NoSQL	Not only SQL
OSM	Open Street Map
PM	Performance Management
QGIS	Free and open-source geographic information system software
RATP	Public transport in Paris
REST	Representational State Transfer

SIRI	European standard for real-time information about public transportation
SNMP	Simple Network Management Protocol
SQL	Structured Query Language
SSH	Secure Shell Protocol
STA	Scheduled time of arrival
STD	Scheduled time of departure
SW	Software
TC	Telecommunication
TRL	Technology Readiness Level
UML	Unified Modelling Language
UMTS	Universal Mobile Telecommunications System
UUID	Universally Unique Identifier
WP	Work Package
XML	Extensible markup language
XSD	XML Schema Definition
ZET	Zagreb Electric Tram
ZIP	An archive file format that supports lossless data compression

0 Executive summary

0.1 Introduction

The EU is taking numerous steps to make the mobility sector more environmentally friendly and efficient, with the aim of reducing emissions and enhancing sustainability. One proposed measure involves the use of smart digital solutions as a practical approach to transforming the transportation sector to achieve these objectives.

Activities within WP2.2, were focused on drafting, developing, testing, and showcasing such smart digital solutions in the form of a software-based decision support system for strategic and comprehensive mobility management (an IT solution for optimizing multimodal traffic).

The proposed solution relies on new technologies and new data sources, incorporating cutting-edge concepts such as big data analysis, data warehouses, AI, and machine learning. The primary beneficiaries of this IT solution are urban agglomerations, public transport operators, airports, and citizens.

0.2 Brief description of the work performed, and results achieved

Throughout the project, the identification of key stakeholders and their requirements formed the foundation for defining the operational scope of the newly developed software solution. Following the identification of user requirements, a new IT system architecture was proposed, aligning with the EU's defined FRAME principles, resulting in a fully operational software solution available for demonstration.

Several use cases were identified in the realm of public transport optimization, airline strategy planning, airport strategic planning, analysis of migration and retention habits of airport users, and transport demand predictions. Analytical artificial intelligence (AI) and machine learning (ML) models were constructed to achieve these goals, drawing on heterogeneous data sets that include common sources (traffic flow data, public transport data, airport data, parking data, etc.) as well as experimental sources (telecom data), which proved to have great potential for mobility analysis.

The final result is a robust IT solution decision support platform, deployed within the closed Zagreb Airport IT system for testing and demonstration purposes. This platform is capable of executing various analytical use cases and is also GDPR compliant. Its success is demonstrated on a number of

use cases in the Zagreb Airport area, as well as on a limited set of use cases in the Paris Airport Charles de Gaulle area.

0.3 Deviation from the original objectives

No significant deviation from the original objectives have occurred during this project. All risks were successfully mitigated, including the challenge of obtaining the telecom data set for Republic of Croatia. However, the data sets available from the Paris Airport Charles de Gaulle were restricted which limited the demonstration capabilities to a smaller number of analytical use cases for that area.

1 Introduction

The transport sector generally has a negative ecological impact and can significantly influence the health and quality of life for everyday people. This issue has been well raised in the European Union through its Sustainable and Smart Mobility Strategy which aims at reducing emissions and increasing sustainability. Additionally, improvements that can be made will also help achieve the goal of environmental efficiency set by The United Nations through the Sustainable Development Goal. Implementing smart digital solutions is a viable option which can help transform the transport sector to meet these goals. A positive effect can be achieved by improving ground transportation in the catchment area of airports. Generally, fostering the use of sustainable and green multimodal solutions for all journeys related to the airport, be it by employees, passengers or supporting services, will help decrease the direct environmental footprint. [1]

Also, airports are in general known for having a significant impact on both the environment and public health and are, therefore, subject of improvement so that they can meet nominated Sustainable Development Goals of the United Nations regarding environmental efficiency. Specific actions are required with the purpose of increasing performance from these perspectives, and one of the most significant ones is related to the improvement of ground transport in catchment areas of the airport. The goal is to foster the use of sustainable and green multimodal solutions for all journeys related to the airport (employees, passengers, supporting services, etc.) to decrease the direct environmental footprint of the airport in the area of ground transportation. [1]

With that in mind, WP2.2. develops an IT solution (platform) for multimodal traffic optimization whose main beneficiaries are urban agglomerations, public transport operators, airports, and citizens. The platform is a software-based decision support system for strategic and comprehensive mobility management which utilizes fused data from different sources:

1. anonymized mobile network originated data about population migration,
2. information concerning public transport,
3. data on passengers and resources from the airport systems.

Acquired data is secured and stored for a limited time abiding by the regulations set by GDPR. Advanced analytics and optimization modules have been developed that use elaborate ML and AI algorithms for traffic predictions. This helps to enhance and significantly improve managing transportation related activities for airports and their gravitational areas. Implementation of this IT solution has the potential to reduce pollution, optimize transport operations, and improve access and multimodal connections to and from airport terminals, aircrafts and at the airport landside.

The IT solution has been developed and tested mostly on the case of Franjo Tuđman Airport Zagreb (in the following text Zagreb airport) and its catchment area, but it can be easily implemented for other airports which are able to supply the needed data input.

2 Solution Overview

The proposed decision support system architecture is presented on Figure 1. It consists of a three-level architecture. The first layer consists of available and required data sources. These data sources are integrated with the second layer—the data aggregation platform using a set of application programming interfaces. Where applicable, standardized protocols should be used to ensure compatibility of the solution. Therefore, transport and public transport data will be supported by using standardized protocols (Datex, Netex, GTFS, SIRI). Other data sources are not covered with standardization, so anonymized telecom data, parking data, airport data and other data sources will be integrated using proprietary data formats.

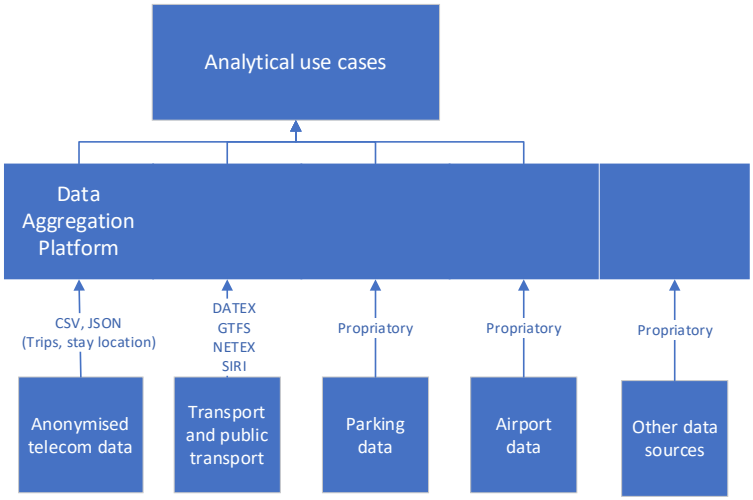

Figure 1 System architecture (high level)

The data is then stored in a data aggregation platform. Conceptualization and realization of the data aggregation platform for urban mobility management has been described in detail in three recent publications [2]–[4]. The platform offers a solution for an intelligent transport system which is applied to traffic management in small to medium sized cities. It includes a three-layer architecture, where the southern interface is used for acquisition of data (input interface), the middle layer is used for operations with data, and the third layer—the northern interface—is used for publication of data toward users/applications (output interface).

In the application layer, applications will be used to perform analytical use cases identified in this project. Besides an analytical engine, an interface will be ensured that enables live streaming of fused data entering the platform. Above these layers, data presentation layers will be established that will enable generation of reports and visual presentation of data on an interactive web GIS tool.

2.1 Data types

Several data sources have been used as input data for decision-making. These data sources include transport data, public transport data, anonymized mobile network data, airline data and other data.

2.1.1 Transport data

Traffic data encompasses information about traffic flow parameters and traffic situations. The Datex standard is employed for the exchange of traffic data among traffic control, information centres, traffic operators, and other stakeholders. This standard enables the collection of data on traffic accidents, ongoing roadwork, and other traffic-related occurrences. The data is formatted in XML and represented by a UML diagram. The message structure includes the publication of measurement location tables and measured data. The data source includes details such as vehicle type and flow rate. Individual vehicle data comprises information about arrival and exit road lanes, vehicle speed, and direction. Additionally, situation data, which includes details about incidents like improper turns, vehicles running red lights, vehicles moving in the wrong direction, and improperly parked or stopped vehicles, is also available. For this study, the focus will be primarily on utilizing vehicle flow data.

2.1.2 Public transport data

Public transport is shared using NeTEx/GTFS or Siri, although the latter is not considered in this research due to the unavailability of data in that format. NeTEx facilitates the exchange of data for passenger information, including stops, routes, timetables, and fares, among various computer systems, along with associated operational data. It enables the collection and integration of data from multiple stakeholders and its reintegration as it evolves across different successive versions. NeTEx categorizes transport modes into scheduled (air, rail, including high-speed and conventional rail, light rail, long-distance coach, ferry, metro, tram, bus, and trolleybus), demand-responsive (shuttle bus, shuttle ferry, taxi, car-sharing, car-pooling, car-hire, bike-sharing, and bike-hire), and personal (car, motorcycle, bicycle). For this study, data is acquired using the NeTEx

protocol from the National Access Point for Multimodal data, focusing on information about public transport lines, stops, and timetables.

2.1.3 Anonymized mobile network data

The traffic sensors commonly deployed in the road network are highly accurate for traditional measurements, such as traffic flow characteristics on specific road sections. However, the demands of road authorities and operators may exceed the capabilities of these sensors or may be focused on road sections lacking sensor infrastructure, leading to the exploration of alternative methods [5]–[9]. One of the extensively researched data sources with substantial potential is the data provided by mobile network operators, which has gained popularity due to the widespread use of mobile phones (market penetration in developed countries exceeds 100%) and the introduction of new generations of mobile networks (3G, 4G, 5G). Each new generation brings notable enhancements in telecommunication (TC) services and consequently results in more valuable accompanying databases generated through network operations. Additionally, each new generation also brings improvements in user positioning, although it is not as precise as Global Navigation Satellite Systems (GNSS), this data remains highly valuable in the field of transport engineering.

The most used data sources from mobile network data include Call/Charging Data Records (CDR), mobile network signalization data logs (CTUM, CTR, etc.), mobile network subscriber database data, and performance counters data.

Call/Charging Data Records (CDR) data is a collection of records documenting user telecommunication activities, as each instance of user communication is logged in this database for future billing purposes. Mobile network signalization data primarily comprises cell trace data, which logs the communication between mobile network terminals and the corresponding antennas. An advantage of these data sources is that they are populated even when the mobile phone is not actively in use (e.g., during a call or data usage). Both of these data sources are user-centric, and it is essential to implement processes, organizational measures, and anonymization techniques to ensure compliance with the General Data Protection Regulation (EU) 2016/679 (GDPR), which governs data protection and privacy in the European Union (EU) and the European Economic Area (EEA). Utilizing this data allows for the anonymized reconstruction of user movements, enabling the extraction of user trips with additional parameters such as origins, destinations, duration, length, and speed. When merged with anonymized user database data (CRM – Customer Relationship Management), supplementary information, such as user gender and age range, can be appended to the extracted trip database. This facilitates various types of analytics for identifying habits, including travel patterns for specific age groups and genders, average speeds, and user compliance with speed limits. Unlike the data sources mentioned above, performance counter data is not user-centric and does not contain data about an individual user. These are regularly collected to provide regular statistical data on the performance of the TC network. LTE/E-UTRA

performance management (PM) counters are aggregated over 15-minute intervals and inherently preserve privacy by representing the number of terminals served by a respective cell, thus functioning as virtual traffic counters.

Telecom data sources offer diverse applications for transport-related analyses. Signalling data, CDR data, and subscriber data can serve strategic and operational purposes, enabling the identification of traffic flow characteristics such as volume and speed, as well as facilitating the analysis of mobile phone usage by drivers to address driving distraction scenarios. Furthermore, these data sources can aid in identifying transport demand by providing origin and destination data. Additionally, performance counters data can be leveraged to create highly efficient operational virtual traffic counters, thereby supplementing locations where traditional traffic counters are not available.

2.1.4 Airport data

Airport data encompasses details about aircraft arrivals and departures, sourced from the airport management system. This data source supplies real-time information on aircraft arrivals and departures, encompassing air carrier details, scheduled and estimated arrival and departure times, stopover information, country of origin, airport details, flight and airline codes, gate or exit designations, aircraft stand information, and status indicators (e.g., arrived or departed). In this study, all this data will be utilized to analyse the correlation between landside transport demand and aircraft arrivals and departures.

2.1.5 Parking data

Parking data provides details about the occupancy and usage of airport parking facilities, sourced from the parking management system. Due to the absence of standardization in this domain, the data is not standardized and is accessible in a proprietary format. The data includes information such as entry and exit times, user types (particularly for taxi vehicles), entry and exit points, and more. This data source will be utilized to identify specific types of airport users, such as taxis, in order to validate the transport mode data acquired from anonymized mobile network data.

2.1.6 Other data

Other data sources might include general and well-known data, such as statistical data on the number of employees at the airport, census data on user migrations, data from transport models, etc. This data will be used to validate results of analyses from different data sources (like anonymized mobile network data).

2.2 Data aggregation platform

2.2.1 Data Management Platform, functional blocks

The aim of this work package was to define requirements on data aggregation platform for its use in urban mobility management within the smart city environment. This research has shown that general purpose data aggregation platform needs to be customized to fulfill application requirements from ITS domain, and to ensure compatibility with defined standard and protocols. These customizations need to be driven by relevant stakeholders responsible for the use of the platform: in urban mobility management, they were involved in the process of definition and prioritization of requirements, and the requirements they had identified were validated by experts responsible for system development. The requirements are grouped as follows: requirements defining data standards and protocols, requirements defining storage, management and permissions, and requirements defining supported transport protocols for data transport. The implementation of proposed requirements enables the general-purpose data aggregation platform to act as a central data aggregation node in ITS applications for urban mobility management. Due to the strong requirement to comply with relevant standards and protocols, interoperability is ensured, and such data aggregation platform can be used in any urban mobility environment following the same standards.

Data aggregation platforms are dynamic, proactive, and heterogeneous systems as well as key enablers for smart city initiatives, targeting the improvement of citizens' quality of life and economic growth [10]. Different smart city applications are based on different architectures, which prevents co-building, convergence, and openness .[11] Data aggregation platform aims to solve the problem of collecting data and ensuring the interoperability of a huge amount of data from heterogenous sources, generated by various stakeholders within the smart city environment.[12] Data aggregation platform in a smart city is used in different business verticals, including energy management, transport, tourism, environment protection, health, home automation, safety and others [11]. In general, one data aggregation platform can fulfil data management and data processing requirements from all these business verticals. However, customizations are required to support use case specific requirements since, for example, use cases in tourism and health domain differ. Besides these

specific requirements, basic functionalities for different use cases are similar or identical, and are in detail covered in the existing literature [13]–[18].

Basic functionalities of the data aggregation platform are transformation, storage and routing of data received at the input—the southern interface—by all relevant data sources, such as senses and/or systems and/or applications, to the northern interface according to the applications and systems that will use this data, i.e. all relevant data consumers. An additional functionality of the platform is that this data flow is performed in a secure way and that access to each portion of ingested and processed data is provided only to the users who need to have the right to access that particular portion of data. This is done using the platform's access control mechanism, which is another general-purpose requirement on a data aggregation platform that is typically domain agnostic (reusable in the same form for multiple domains). For this project, a general-purpose data aggregation platform has been chosen for customization. The chosen platform has an internal platform attribute-based access control mechanism based on finely grained permissions that are granted to various platform users on a need-to-have basis. This platform has been developed by Ericsson Nikola Tesla [19]. Architecture of the platform is designed in a modular way in line with modern software development principles to allow easy scalability in terms of new future functionalities or capacities. In conceptual architecture the platform can be shown as in Figure 2.

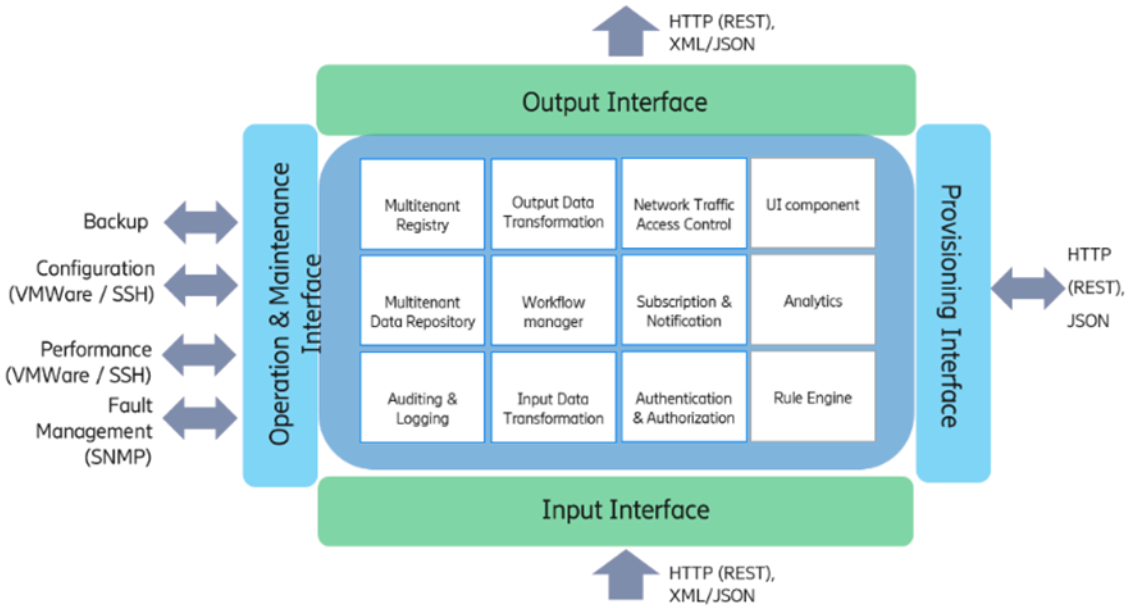

Figure 2 Conceptual architecture of the data aggregation platform, based on [19]

General data aggregation platform includes a three-layer architecture, where the southern interface is used for acquisition of data (input interface), the middle layer is used for operations with data, and the third layer — the northern interface — is used for distribution/publication of data toward users/applications (output interface). Additionally, data aggregation platform uses an information model registry to store data on device specifications and instances, information on solution owners, domains, organizations and users, and information and credentials for data producers and data consumers such as applications. Data Repository enables data storage with encryption support. The data storage is independent of the data source's (e.g. device's) protocol and data format. Platform enables publisher-based subscriptions and notifications. One of the features is transformation of data stored or passing through the Platform or transformation of proprietary data into data-source--independent data. Additional features include support for workflows (application-specific workflow execution with configurable workflow engine), security (fine-grained and customizable security, unified role-based and permission-based access control for all traffic and provisioning interfaces), rule engine (support for building simple or complex business rules), analytics (streaming calculations and aggregations), configuration, logging, etc. However, this platform must be customized to be used for the ITS application based on the application and/or use case specific requirements, and these requirements are identified and defined in the remainder of this chapter.

Requirements specific to the application of the platform in Intelligent Transport Systems have been defined. The first group of requirements refers to data standards and protocols. The platform should have support for the exchange of "basic general traffic information related to road safety" in DATEX II format (CEN / TS 16157) as defined by EC Delegated Regulation No. 886/2013 and the Datex II 2.3 domain. The platform should be able to process the data defined in package 5 of the Datex specification entitled "Measured and Elaborated Data Publications" in the "Elaborated data" segment. The processing refers to data on travel times, status, traffic values and data on weather values. The platform should enable the exchange of information on multimodal travel, namely "static travel and traffic data" using the NeTEx standard (CEN / TS 16614) in accordance with the Commission Delegated Regulation (EU) 2017/1926. The implementation of the following parts of the wider NeTEx standard are needed on the platform: Part 1. Fixed network topology and Part 2. Timetables. Implementation of Part 3. Fare data is not planned. The platform should also enable the exchange of information on multimodal travel, namely "dynamic travel and traffic data" using the SIRI standard (CEN / TS 15531) in accordance with the Commission Delegated Regulation (EU) 2017/1926. As part of the OLGA project, implementation of the vehicle information service —SIRI VM— "The Vehicle Monitoring Service" is needed.

The following set of requirements refers to data storage, management, and permissions. Therefore, the platform should be able to store data (storage of received data in original, unprocessed form and storage of aggregated data). For each recipient of distributed data, the platform must be able to define a subset of the received data to be distributed to that recipient. For each recipient of distributed data, the platform must be able to define access rights to the appropriate subset of received data. Any data received will be distributed only to those recipients who have a defined need and right of access to that data. The platform should be able

to distribute data to third parties through appropriate communication and application interfaces, namely the retrieval of historical data through the REST interface (the so-called data pull mechanism) and the distribution of subsets of received data to interested and authorized recipients immediately upon receipt (the so-called data push mechanism). The platform must ensure that the interface for receiving and distributing data extends to formats other than those explicitly specified within this project. Several demonstration acceptance formats will be implemented through this project, while additional ones will be able to be reported and the platform expanded later. Example data includes data on the number of passengers in public transport, receiving information on the condition of traffic light equipment, and crowdsourcing data. The platform must be able to distribute data in a different (transformed) format than the one in which it accepted data at the input. The platform should also have the possibility of enriching the data it accepted at the entrance with additional contextual data from the information model by which the access point was provisioned in advance (before the acceptance itself). It should be possible to use the data that enriched the input data when defining the distribution (output) data format. The platform must enable the addition (provisioning) of new data sources, change of existing ones, as well as deactivation of existing sources. The platform should be able to manage data sources and outputs in a way that ensures access rights to data sources in such a way that access is provided to those users who have the right of access, as well as granulation to access the entire set or part of data.

The final set of requirements is defining supported transport protocols for data transfer. The platform should be able to receive and distribute traffic data using the RESTful (Representational state transfer) protocol as a transport protocol. The platform should also be able to provide for the information model using the RESTful (Representational state transfer) protocol. The information model is used to represent different real-world entities related to the data sources and recipients needed for the access point to function properly in terms of security, enrichment, data source management, and data distribution described in other access point requirements. The access point must be able to receive and distribute traffic data using the MQTT (MQ Telemetry Transport) protocol (ISO / IEC PRF 20922) as a transport protocol.

2.2.2 Traffic data flow – simplified view

Traffic data flow in simplified view is presented in Figure 3.

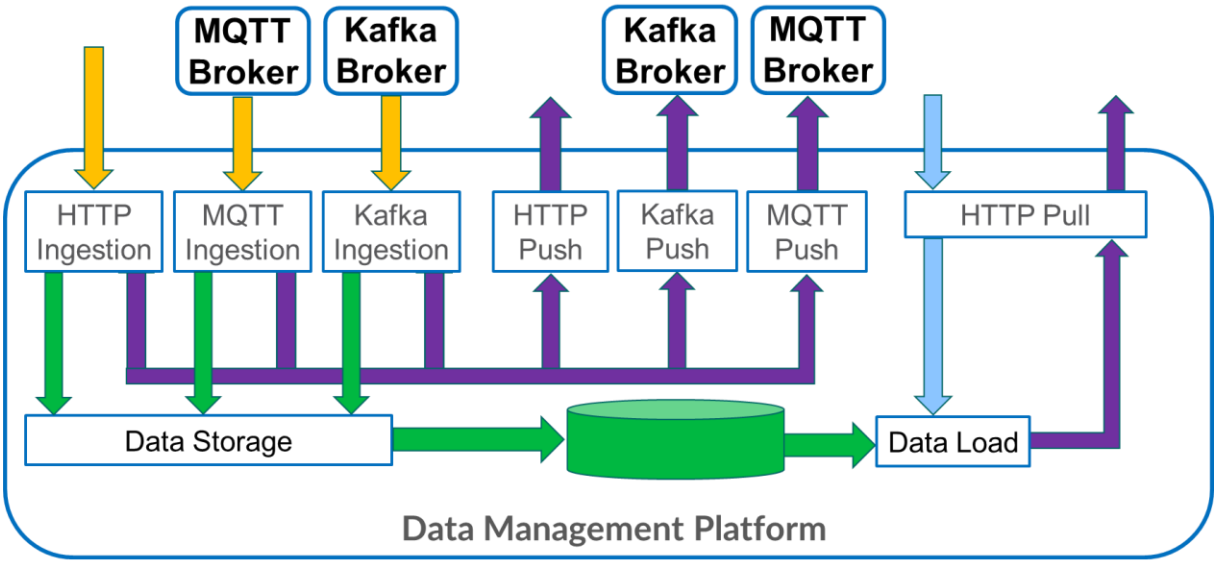

Figure 3 Data Management Platform traffic data flow.

Data Management Platform can receive or deliver data in several ways:

- via HTTP protocol and it's REST API. For this type of data ingestion there are various input adapters which support JSON, Simplified JSON, XML and LwM2M messages but can be (and in this project are) extended to use case and domain specific formats, such as DatExII, NeTEx and SIRI.
- via MQTT broker. Our MQTT client subscribes to predefined topics and after that MQTT broker pushes new data to the platform's MQTT client that ingests the message into Data Management Platform to save it into data storage. The formats
- via Kafka broker. Firstly, Data Management Platform's Kafka client subscribes to Kafka topic on Kafka broker. After that Kafka broker sends new data to Kafka client that is subscribed on topics that have new data. Lastly, Kafka client is forwarding message to Data Management Platform to save it into data storage.

All ingested data is validated, enriched authorized and processed in custom use case specific ways (if needed) before being saved in the platforms internal data storage as well as delivered to the subscription processing mechanism to be provided real-time to the subscribed data consumers. The platform's data storage uses a hybrid database model which consists of two databases:

- Postgres – for provisioning data (hierarchy of devices, its sensors and resources)
- Apache Cassandra – for traffic data (actual payloads of all the devices, its sensors and resources)

Data Management Platform supports subscription. Data consumers can subscribe to various cross-sections of the ingested data using subscriptions that can filter the data based on its enriched context. The subscribed data consumers can receive their filtered data through HTTP Push, Kafka Push or MQTT Push mechanism.

Besides retrieving data via push mechanism, the stored historical data can also be filtered and retrieved using a classic HTTP GET request (HTTP Pull).

2.2.2.1 Traffic data flow – detailed view

Traffic data flow in simplified view is presented in Figure 4.

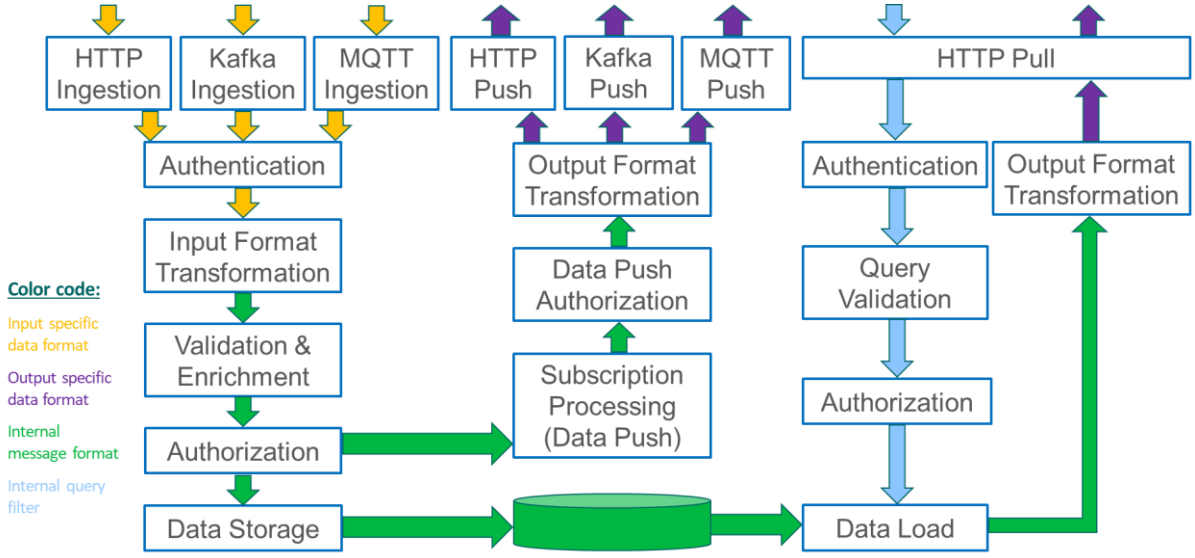

Figure 4 Data Management Platform traffic flow with internal view and formats

Figure 4 shows a detailed diagram of traffic flow with internal view and formats. All input requests (HTTP ingestion, Kafka ingestion, MQTT ingestion) must go through several stages:

- Authentication – To be able to ingest or to retrieve data, user must authenticate itself with correct username and password. For HTTP ingestion, there are two supported authentication types: digest and basic authentication. If the sender fails the authentication process, the message is discarded, and the proper failure response is returned to the data producer (sender of the data).
- Input Format Transformation – Data received in JSON, Simplified JSON or XML format, as well as any use case and domain specific format, must be converted into the internal format of the Data Management Platform.
- Validation & Enrichment – In this step the message is checked for consistency and its content is validated against the information model provided in the Data Management Platform. The validation step verifies that all entities listed in the input message exist in the Data Management Platform and

that they are properly related. Once the validity of the message is confirmed, the entities that might be missing in the originally received message are filled in from the information model provisioned in the Data Management Platform.

- Authorization – this process checks that the data producer (sender) is authorized to save the message content into the Data Management Platform and to forward it to the destinations specified in the message (if any such destination is specified in the message). Upon successful authorization, the message sender is informed that the message is successfully processed and further Data Management Platform actions on the message are performed asynchronously.
- Data Storage – In this step Data Management Platform stores the message into the database. This can be done synchronously (which is the default) or as part of the asynchronous processing.

Additionally, after the message is passed through the authorization phase, Data Management Platform checks if there are subscriptions for the message data, and then checks if subscribers are authorized to receive all or parts of the data contained in the received message. If the data matches one or many subscriptions and the subscribers are properly authorized, the message is transformed into the format in which the subscriber wants to receive the data and sent as a HTTP POST request and/or via Kafka and/or MQTT Push mechanisms.

2.2.3 Platform Information Model

To be able to story any type of data into Data Management Platform, there are some prerequisites to be met in the first place. Platform has its own information model which consists of several groups of entities shown in Figure 5 Platform Information Model:

- Access Control – hierarchy of entities that consists of principal, accounts connected to principal, roles connected to accounts and roles' permissions. Complete hierarchy is used to precisely define which account has access to specific data in the Platform.
- User and Organization Hierarchy – consists of operator entity, domain applications connected to it, enterprise customers connected to domain applications or to other enterprise customers, and lastly, users connected to enterprise customers.
- Data Source Type and Instance Hierarchy – In this hierarchy there are two parts: hierarchy of specifications and hierarchy of their instances. Every specification has its corresponding instances (it can be anything from zero to any number of instances). Top level of specification hierarchy is device gateway group specification, which has device gateway specifications connected to it. Next in the hierarchy are sensor specifications – they are connected to their device gateway specifications. Lastly, there are resource specifications, connected to their sensor specifications. Specifications hierarchy can be seen as a blueprint for hierarchy of instances. Actual data is saved into Platform under instances hierarchy. Specifications hierarchy helps when the user wants to retrieve all the data under some specific entity (e.g. some device gateway group specification or any other part of specification hierarchy).

- Subscription Mechanism – consists of two entities: subscription filter and addressing info. Subscription filter describes a cross-section of the data that the subscriber is interested in, while addressing info connected to it describes the destination address (where the data needs to be pushed) and format of the outgoing (push) message.

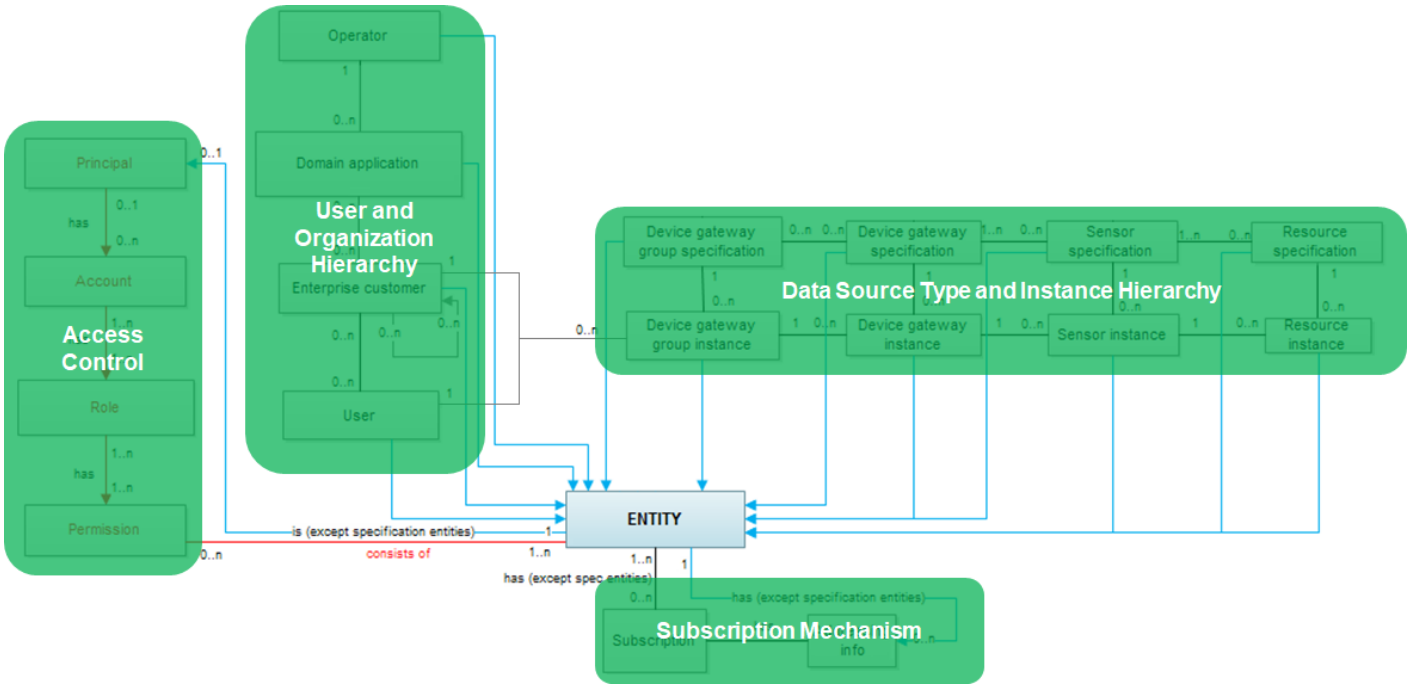

Figure 5 Platform Information Model

2.2.3.1 Internal Traffic Message Format

Before storing the data received on the endpoint for data ingestion, the message must be converted to internal message format described in Figure 6 Internal Traffic Message Format.

Internal message can consist of multiple data elements. Each element has timestamp (date and time when data is received), actual value (data) that needs to be saved, source address to which the data belongs to and additionally it can have metadata in a form of multiple key-value pairs.

Message is saved under specific user, data source type and instance hierarchy. That's source address of the saved data.

When retrieving data, address of the data must be sent. It doesn't need to be precise to the bottom level of hierarchy (i.e. resource specification or instance), but instead it can consist just of one of the top-level entities in the hierarchy (e.g. device gateway group specification). Of course, when such partial address is given, this means that the Platform will fetch all the data that are found under the specified address (e.g. all the data under the specified device gateway group specification – that's data from all the gateway instances, sensor instances

and resource instances attached to all gateway group instances which are attached to specified gateway group specification).

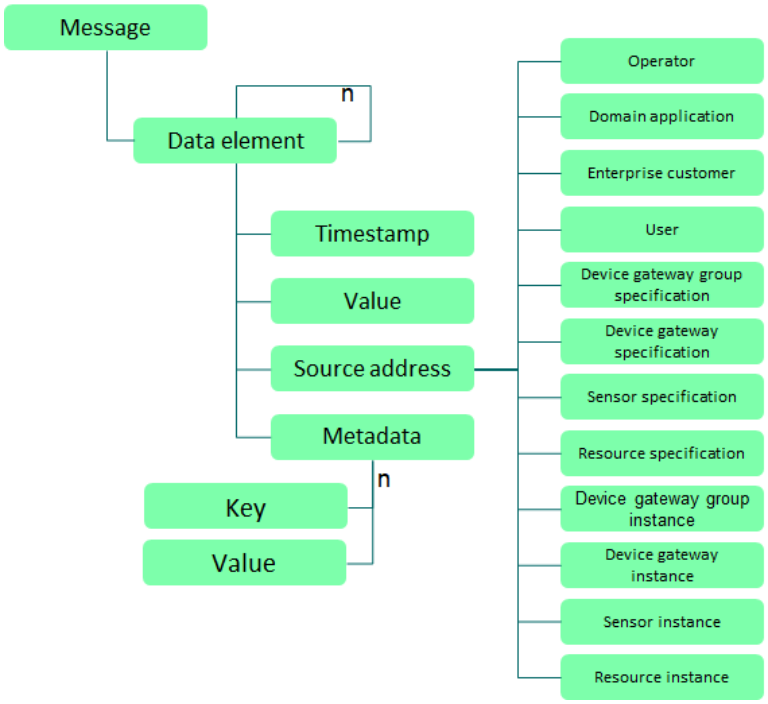

Figure 6 Internal Traffic Message Format

2.2.4 Deployment

To be able to process a lot of requests and to be fault tolerant, there are some recommendations how to deploy the Data Management Platform used in this solution. Figure 7 shows an example deployment with eight application nodes, one PostgreSQL node and one Apache Cassandra cluster that consists of four nodes. To have high availability, there should be two PostgreSQL nodes with automatic replication between them.

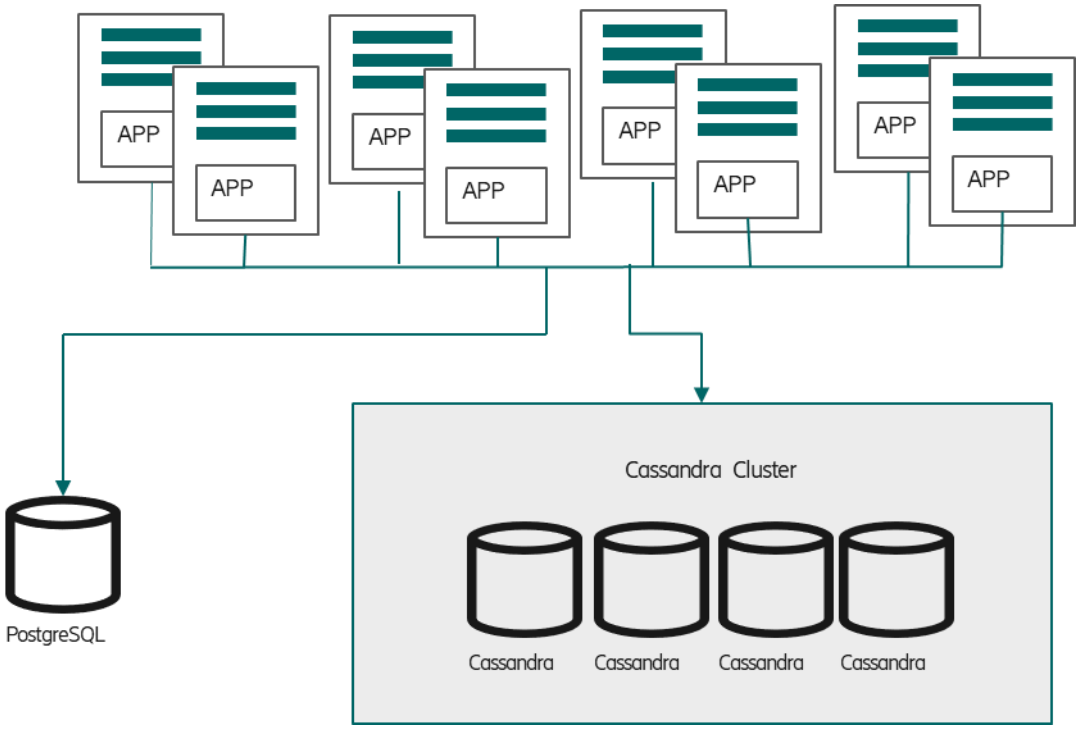

Figure 7 Data Management Platform deployment.

Data Management Platform can be deployed in either virtualized (including cloud) or on physical servers.

High availability is supported by design along with scalability. This means that for use cases where Data Management Platform needs to handle very high number of requests it's possible to deploy more instances of application, PostgreSQL and Cassandra servers.

Minimal configuration consists of the following:

- Application server
- PostgreSQL database
- Apache Cassandra database

High availability configuration:

- Application server (more than 1 node, linearly scalable)
- PostgreSQL database (2 nodes)
- Apache Cassandra database (minimum 3 nodes, linearly scalable)

CPU, RAM and HDD dimensioning depend on traffic interaction model.

Application server has installed Apache Karaf, a so called modulith runtime. This means that everything inside Karaf is built as modules. Karaf provides several frameworks and different kind of applications: REST/API, Web, Spring Boot, and much more.

Key features of Karaf Runtime:

- Hot deployment
- Dynamic configuration
- Logging system
- Provisioning
- Shell console
- Remote management
- WebConsole
- Security
- Instances management
- Docker & Cloud ready

Data Management Platform consists of multiple modules running inside Apache Karaf Runtime.

PostgreSQL database is a relational database that the platform uses for storing data about provisioned user and device hierarchy. It's open source and very popular in the IT community. It has a very large community, and it has been actively developed for more than 35 years. This ensures it has a very good support with frequent security updates coming out. In case of new serious security vulnerabilities, a new patch will be available very soon which is very important.

Apache Cassandra is an open-source NoSQL distributed database. It supports scalability and high availability out of the box. Also, it is fault tolerant and independent on the used hardware. This makes it perfect for storing mission-critical data. Cassandra is deployed as a cluster of multiple Cassandra nodes. Minimal high availability configuration consists of 3 Cassandra nodes. This way Cassandra also supports fault tolerance.

Distribution of data to different nodes provides power and resilience. Data from the whole database is distributed in parts on every node. This is done through partitions defined with partition keys in the table. That way, if the table has a lot of data, which is almost always the case in NoSQL databases, and we want to retrieve just a part of the data from the table or to run some additional operations on this data, every node does just a part of the job in parallel. This makes the whole process very fast.

One more very powerful feature of Apache Cassandra is replication. Instead of distributing the data so that every node has a different data set, a part of data is made redundant (it is available on several nodes). On how many nodes the same data is available is defined by a replication factor. If the replication factor (RF) is 3, this means that one portion of data is replicated to total of 3 (replica) nodes, ensuring reliability and fault tolerance. This means that we have total of 3 nodes with the same data. An example can be seen on Figure 8. If we set

RF to 1 that would mean that one portion of data is not available on multiple nodes, but just on one. Because of that if one node fails, that part of data becomes unavailable, which should be avoided.

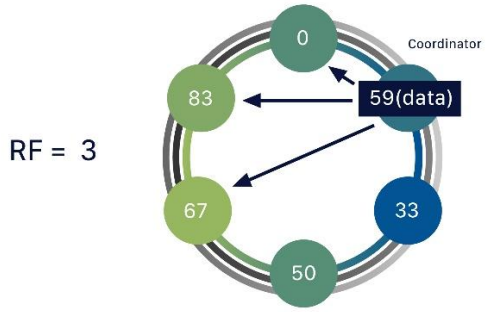

Figure 8 Replication diagram with 6 Cassandra nodes[1]

Replication ensures reliability and fault tolerance. Because of this feature there is no problem if one or even several nodes go offline because lost node will be synchronized with other nodes after it comes back online.

It must be noted that there are two types of nodes: coordinator and regular node. The coordinator node isn't a single location. It's simply the node that gets the request at that moment. This means that any node can act as the coordinator.

Another great Cassandra feature is consistency level. This represents the minimum number of Cassandra nodes that must acknowledge a read or write operation to the coordinator node before the operation is successful. Generally, consistency level is set based on replication factor.

Figure 9 shows an example of 6 Cassandra nodes, with replication factor (RF) set to 3 and consistency level (CL) set to quorum. Quorum is referring to majority. In this case that's 2 replicas or RF/2 + 1, therefore the coordinator node will need to get acknowledgement back from two of the replicas for the query to be considered as a success.

[1] https://cassandra.apache.org/_/cassandra-basics.html [12.09.2024]

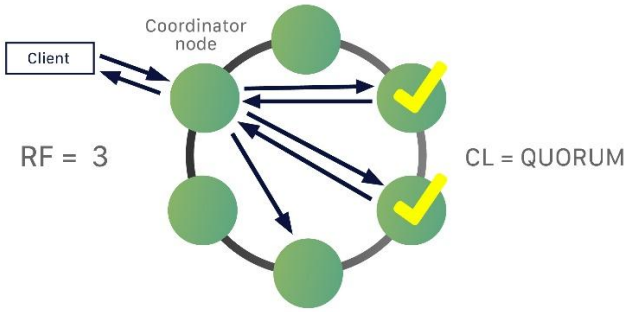

Figure 9 Six Cassandra nodes with RF = 3 and CL = quorum[2]

This means that if one node is down, everything will be fine like nothing happened but if two nodes are down, query will not be considered a success.

2.3 Analytical use cases

The goal of the project was to develop a decision support solution for strategic planning and management of activities related to the land transportation within the catchment area of the airport. The solution aims at enhancing, optimizing, and significantly improving available transport options. For this reason, relevant stakeholders were involved to help identify and target areas which can be improved. All proposed measures fall within the category of Intelligent Transport Systems (ITS) so the chosen methodology for use case development is a combination of the Delphi method and utilization of the FrameNext ITS architecture (FRAME). An ITS architecture is a conceptual design that defines the structure and/or behaviour of an integrated ITS. The Delphi method surveys a panel of experts in a process to reach a group opinion or decision [20]. Almost all top-level requirements and functionalities or use cases of ITS applications and services considered for implementation in the European Union are comprised within the FRAME Architecture. The architecture is relevant enough that it is being used as a reference by all ITS architects and has the basis to be the foundation upon which other types of architectures can be built. This will enable compliance at interfaces between other systems and provide seamless service to cross-border travellers, additionally an open European market of compatible components can be established [21][22]. Therefore, the Delphi methodology has been used to extract knowledge from relevant experts nominated from key stakeholders, while the FRAME architecture was used to shape the user requirements in the matter to fit the standardized requirements. During a series of workshops, a number of key stakeholders (12) were interviewed so as to identify key issues of landside transportation in the catchment areas of the airport. Relevant stakeholders involved

[2] https://cassandra.apache.org/_/cassandra-basics.html [12.09.2024]

representatives of neighbouring cities, representatives of the airport and connected companies, representatives of public transport companies operating in the area. They were nominated in order to fulfil the FRAME next principles, according to which, all relevant stakeholders' groups have to be represented (stakeholders that want ITS, use ITS, rule ITS, make ITS and service providers).

Five workshops with the relevant stakeholders were organized. This resulted in identifying analytical use cases and proposing land transport optimization in the narrower airport area. User aspirations are addressed with the analytical use cases, while user needs are addressed after the research team has proposed a research methodology, required data sources and infrastructure, and has consolidated them in a consistent manner. The consolidated requirements where then presented to stakeholders, and after a common agreement has been achieved, use cases have been prioritized and completed. These use cases are set up as key requirements for the decision-making system. The use cases are the following:

1. *Public transport optimization for nearby residents and airport users (both passengers and employees).*
2. *Strategic planning of the airport gravitational areas and catchment zones.*
3. *Airline strategic planning*
4. *Analytics of migration and retention habits of passengers.*
5. *Transport demand prediction*

Use cases will be described in the following chapters.

2.3.1 Public transport optimization for nearby residents and airport users (both passengers and employees).

The goal of this use case is to optimize existing public transport lines and/or introduce new public transport lines in the Zagreb airport catchment area which serve residents, passengers, and airport employees.

Benefit: Public transport optimisation and shift towards sustainable landside transportation
Beneficiary: Public transport operator
Indirect Beneficiary: Citizens, City/Municipality
Decision level: Strategic/Operational
Related KPI-s: MS1.1, MS1.2, SOC2

2.3.2 Strategic planning of the airport gravitational areas and catchment zones.

Analysis of the users' migrations from mobile network data. Trips that have their origins and/or destinations at the airport are considered. Goal is to identify the real catchment areas, enhance sustainability of the transport connections in those areas, and further analyse potential areas which are not currently served by the airport but are expected to be.

Benefit: Increasing the competitiveness of airport by identification of real gravitational/catchment areas, enhancement of sustainable transport connections for those areas, analysis of potential areas which are not, but are expected to be served by analysed airport
Beneficiary: Airport
Indirect Beneficiary: Citizens
Decision level: Strategic/Operational
Related KPI-s: SOC2, SOC3, GHG2

2.3.3 Airline strategic planning

Insights on travel demand are gathered for planning of airline operations based on information of the initial country of origin and destination of airline users.

Benefit: Identify countries not initially served by airline companies, with significant travel demand.
Beneficiary: Airport, Airlines
Indirect Beneficiary: Citizens
Decision level: Strategic/Operational
Related KPI-s: No KPI's directly related

2.3.4 Analytics of migration and retention habits of passengers.

Based on the analysis of migration patterns and retention habits of passengers, distinction between passenger types (tourist, business users) can be made along with optimization of service/offering for targeted users.

Benefit: Analysis of migration and retention habits of passengers, identification of passenger type based on its migration patterns (tourists, business users). Optimization of service/offering (accommodation, transport, leisure/business offerings) for targeted passengers types.
Beneficiary: Passengers
Indirect Beneficiary: "Airport, Local enterprises, City/Municipality
Decision level: Strategic/Operational
Related KPI-s: MS1.1, MS1.2

2.3.5 Transport demand prediction

Analysis of transport demand for future predictions based on information on airport arrivals and departures, and utilization of public transport services serving the airport area, including transport on demand providers. Based on historical data, migration patterns of airport passengers following the arrival of an aircraft (e.g. the number of passengers leaving airport with a provider of on-demand transport or public transport), the demand will be predicted for a specific transport mode/service (e.g. on-demand transport, rent-a-car, etc.)

Benefit: Predicted demand for specific transport mode/service (i.e. transport on demand, rent-a car...) based on information on aircraft arrival by criteria such as origin, type (regular, charter, domestic, international...), occupancy or other criteria
Beneficiary: Transport services providers
Indirect Beneficiary: Passengers
Decision level: Tactical
Related KPI-s: GHG2, MS1.1, MS1.2

Identified analytical use cases will be implemented as a part of the newly developed IT Solution (platform) for traffic flow optimisation and the results will be demonstrated for Zagreb Airport area (all use cases) and Paris Airport area (limited number of use cases). This tool will demonstrate its general applicability, since it takes into consideration local data sources and local specificities. Due to its modularity and flexibility, tool will support further use cases relaying on similar data sets where the main goal is support of decision making based on objective measurements.

3 Implementation

3.1 Data preparation and ingestion

3.1.1 Transport (Traffic) data

Transport (traffic data) has been implemented using Datex II format. DATEX II is the reference data standard in the European Union for road traffic and travel information. DATEX II is the electronic language used in Europe for the exchange of traffic information and traffic data. It has grown from a standard for the exchange of traffic related data between road-traffic control centres to a coherent set of standards supporting the digitalisation and automation of the entire road transport ecosystem to contribute to a safe, green and efficient travelling of persons and goods. DATEX II supports two closely related use cases: The provision of information from a data source to a data consumer on one hand, and the support of joint traffic management operations of collaboration competent authorities on the other. DATEX II enables the standardised data-provision from the lightest vehicle types like cycling up to the information relevant for heavy lorry convoys. It enables to digitally express in a standardised way the dynamics of road transport infrastructure availability and incidents happening on it. Different parts support the historical, actual and prognosed usage combined with the conditions of the networks and the interventions/temporary changes of the competent authorities in this domain. Road transport infrastructure covers everything relevant for the end user to manage his road usage and therefore varies from the road/street itself, the applicable conditions of use like traffic management measures, UVARs roadworks etc, to parking, to charging and refuelling and all that is directly related to that.

"Point by coordinate ETRS89 (or WGS84)" will be used for referencing the location of the measuring point, i.e. display using geographic latitude and longitude.

In addition to level A *groupOfLocations*, level B extensions *locationInfo* and *roadInfo* have been added. This enables road names, road markings, sections, operators, location names, countries, regions along with the coordinates to be added (Table 1).

Table 1 Location reference

Enumerated value name	Designation	Definition
locationCountry	Location country	Country where the event lies. The info is important.
locationRegion	Location region	Specification of the federal state, so that messages can be filtered by region. It can also affect several states.
locationName	Location name	If the event location is not on any street. Eg: on a POI.
locationText	Free text for the location	eg. A23 Klagenfurt West, St.Veit at the Glan city center.
roadName	Name of the road	Which the linear element forms a part.
roadNumber	Identifier/number of the road.	Which the linear element forms a part.
roadOperator	Name of the road operator	Name of the responsible road operator for this road
roadSection	Road section	Specification of the road sections

There is a need to expand the model with additional data about the vehicle, and the following attributes have been added: id, speed, direction of movement, traffic lane, event name, which are displayed in the event expansion graph (Figure 10). The DATEX II model becomes level B by extension and is still readable by level A.

Extension of model „Traffic data“:

- *arrivalRoadLane*
- *exitRoadLane*
- *minVehicleSpeed*
- *maxVehicleSpeed*
- *deltaVehicleSpeed*
- *arrivalVehicleDirection*
- *exitVehicleDirection*
- *vehicleRegistrationPlateIdentifier*

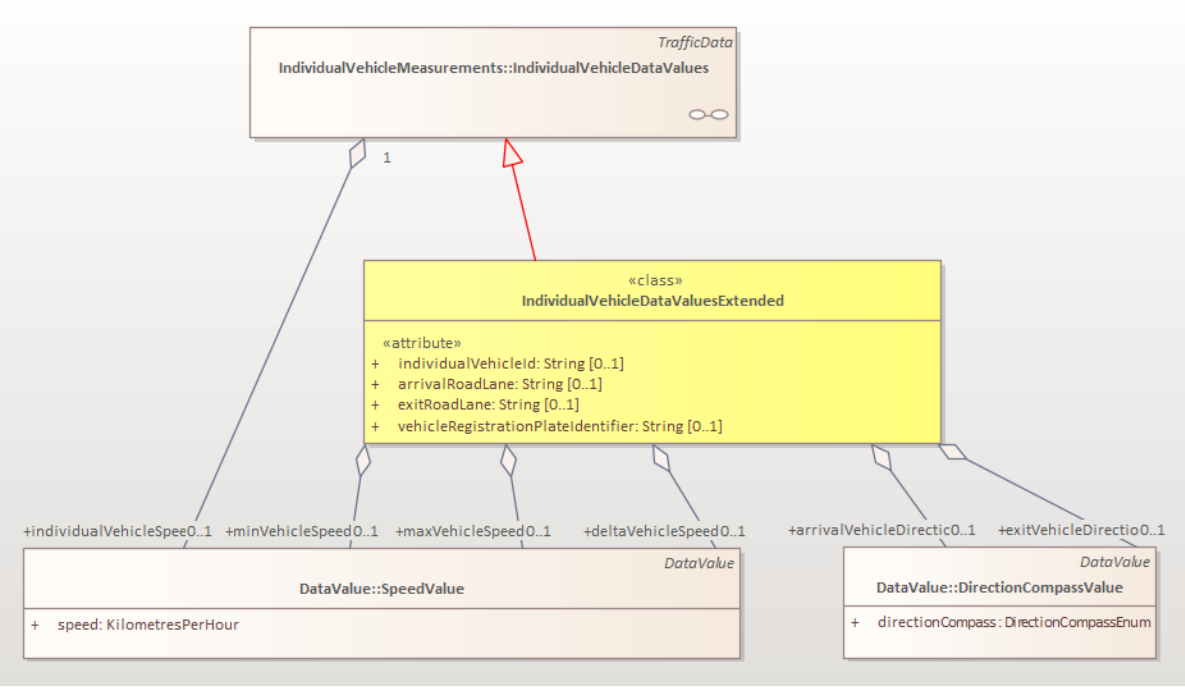

Figure 10 Traffic data profile extension

The Datex II model was created by XMI in Enterprise Architect, then the selection of the DATEX II model was made in the D2 application, the necessary classes (Figure 11) were selected and the XSD file was created. The test XML files were tested with the xml validator: https://www.freeformatter.com/xml-validator-xsd.html.

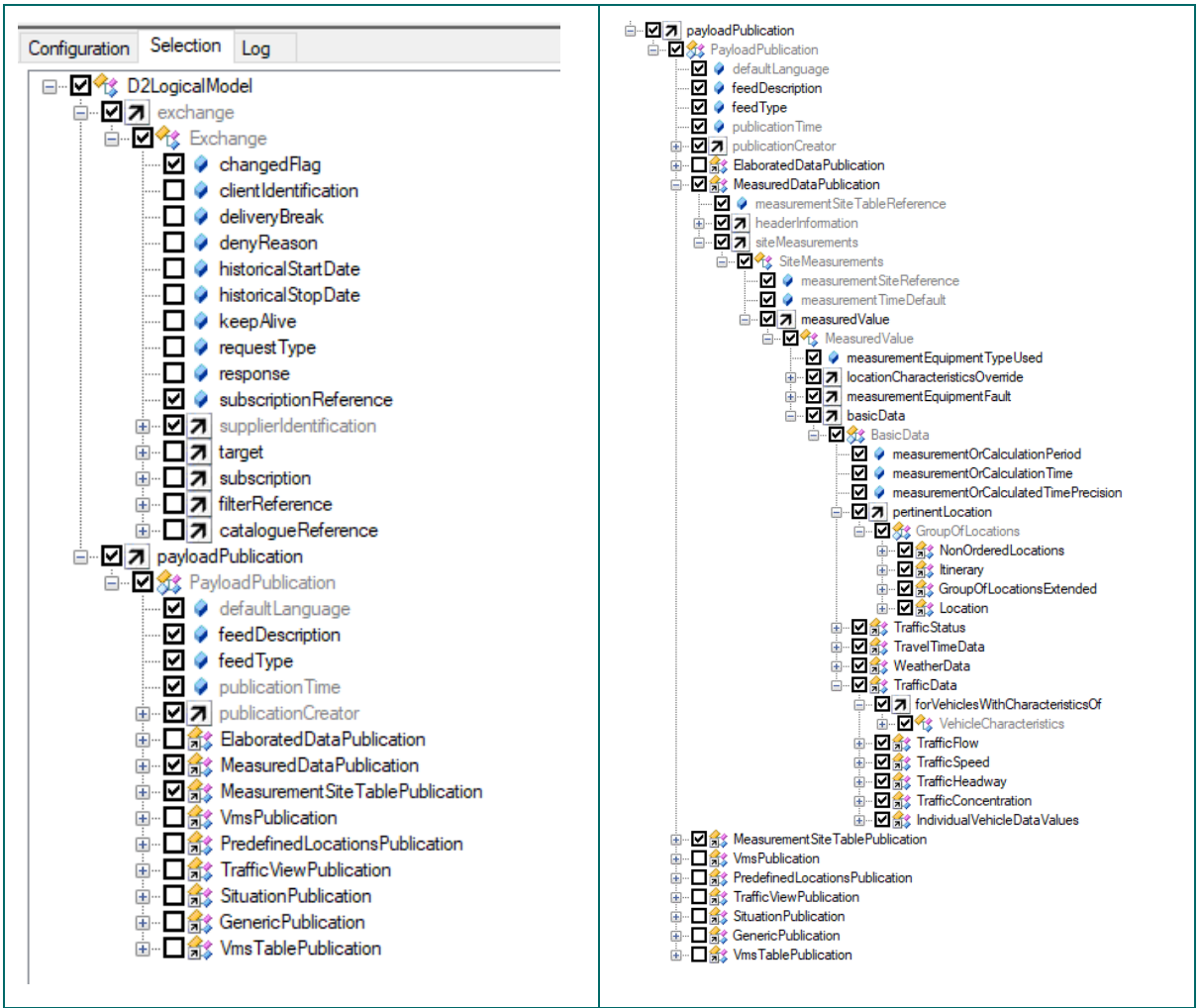

Figure 11 Image of selected classes in the extended DATEX II Traffic data profile model

Examples of measurement values (Table 2) for the use case of counting and detecting the passage of vehicles through an intersection or a specific road zone.

Table 2 Example of measurement values for the user case of counting and detecting the passage of vehicles through an intersection or a specific road zone.

Atribut	Tip	Vrijednost
country	String	hr
nationalIdentifier	String	HRV
payloadPublication	String	**MeasuredDataPublication**

publicationTime	Date Time	2024-08-27T12:06:22.4181616+02:00
MeasurementSiteTable id	String	GUID080cb43b-90aa-4508-aa1e-8ff073ef8efd
confidentiality	String	internalUse
informationStatus	String	test
MeasurementSiteRecord id	String	326290386
measurementTimeDefault	Date Time	2020-08-27T10:46:42.09+02:00
type	String	IndividualVehicleDataValues
measurementOrCalculationTime	Date Time	2024-08-27T10:46:42.09+02:00
measurementOrCalculationPeriod	String	240
pertinentLocation type	String	Point
roadNumber	String	D8
roadOperator	String	Croatian Roads
locationForDisplay latitude	Numeric	45.3270271
locationForDisplay longitude	Numeric	14.4475963
vehicleType	String	Car
Speed	String	50
arrivalTime	Date Time	2024-08-27T10:46:42.09+02:00
exitTime	Date Time	2024-08-27T10:46:42.09+02:00
individualVehicleId	String	25415
arrivalRoadLane	String	3
urlLinkDescription	String	Slika nepropisno zaustavljenog vozila
arrivalVehicleDirection	String	east
exitVehicleDirection	String	east
vehicleRegistrationPlateIdentifier	String	RI342FV

3.1.2 Public transport data

The General Transit Feed Specification (GTFS) defines a common format for public transport schedules and associated geographic information. GTFS feed is composed of a series of text files collected in a ZIP file. Each

file models a particular aspect of transit information: stops, routes, trips and other. Used GTFS data were needed to locate public transport vehicles.

Table 3 Schema of GTFS data.

GTFS data		
Atribut	type	Definition
Stop_id	String	Stop id has unique string combination for every route
Stop_name	String	The name of the station
Stop_lat	Float	Latitude of stop_id
Stop_lon	Float	Longitude of stop_id

3.1.2.1 GTFS Data

The General Transit Feed Specification (GTFS) is an Open Standard used to distribute relevant information about transit systems to riders. It allows public transit agencies to publish their transit data in a format that can be consumed by a wide variety of software applications.

GTFS consists of two main parts: GTFS Schedule and GTFS Realtime.

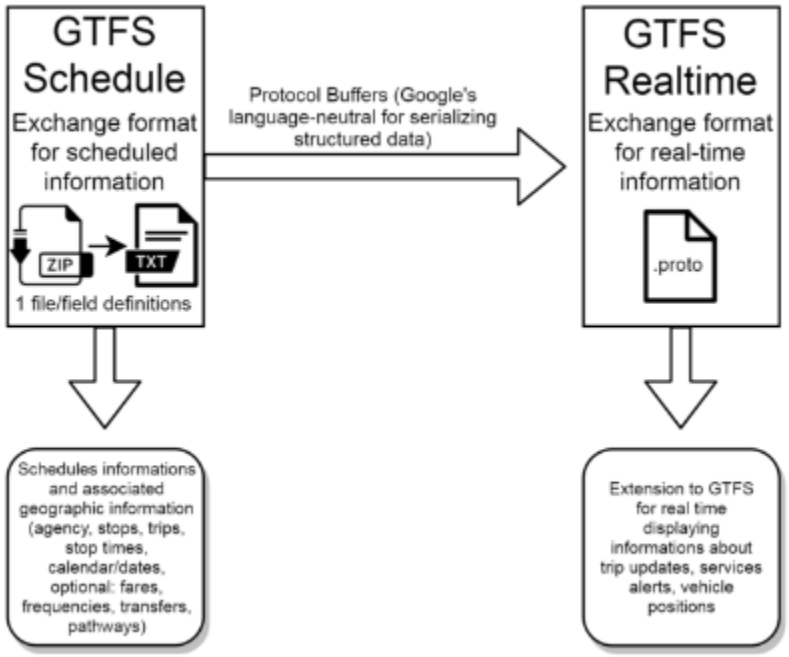

Figure 12 GTFS Schedule and GTFS Realtime File distribution.

3.1.2.2 GTFS Schedule

GTFS Schedule is a feed specification that defines a common format for static public transport information. It is composed of a collection of simple files, mostly text files (.txt) that are contained in a single ZIP file.

Each file describes a particular aspect of transit information such as stops, routes, trips, etc. At its most basic form, a GTFS Schedule dataset is composed of 7 files: agency.txt, routes.txt, trips.txt, stops.txt, stop_times.txt, calendar.txt and calendar_dates.txt.

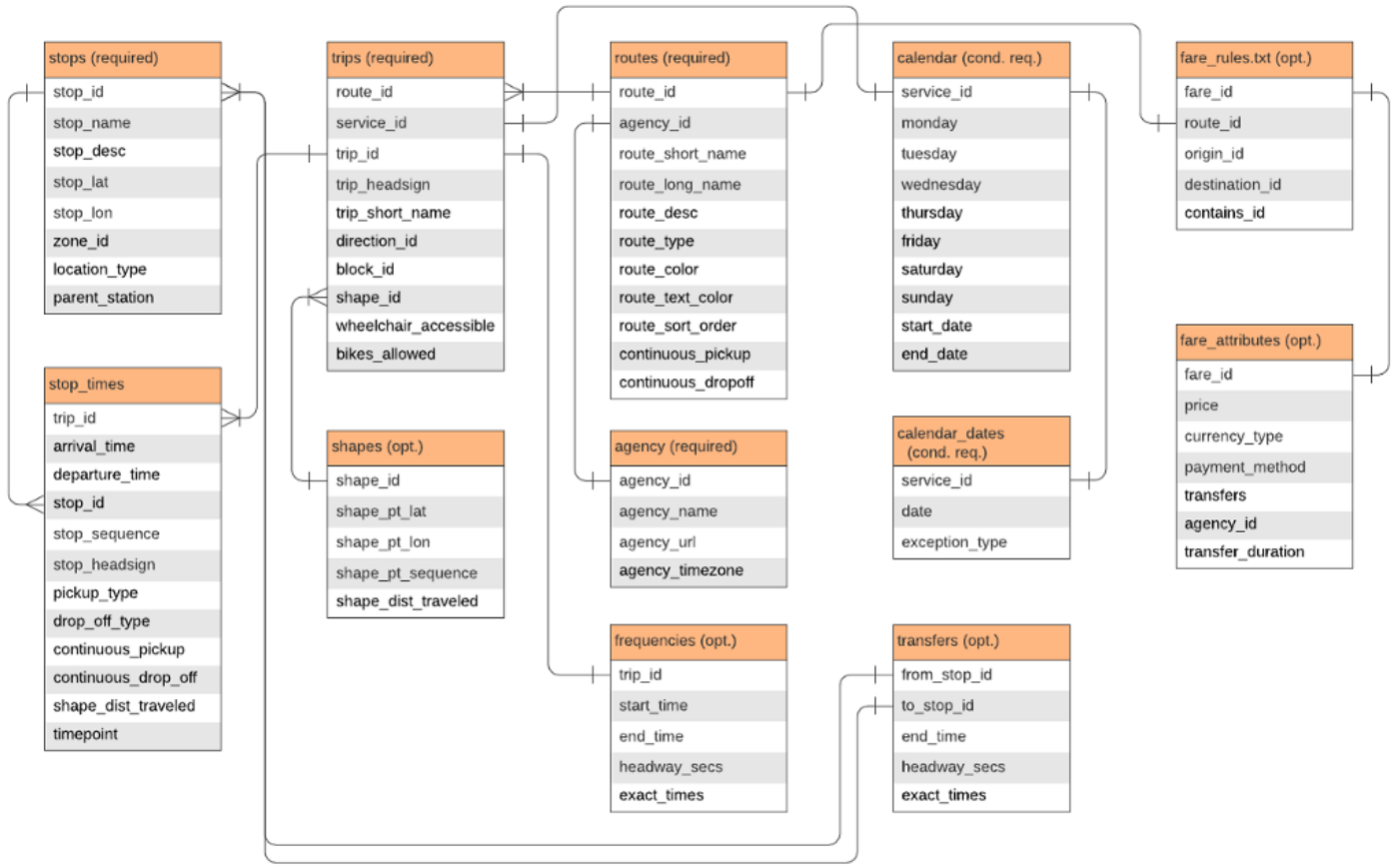

Figure 13 GTFS Schedule Structure.

Along with this basic set of files, additional (optional) files can also be grouped to provide information about other service elements, such as fares, translations, transfers, in-station pathways, etc. Currently there are more than 15 optional files that extend the basic elements of GTFS, including locations.geojson which introduced a new format besides text files (.txt) which can be used to represent geographical areas.

The source of truth for all GTFS Schedule files is the official GTFS Schedule Reference[3], which provides detailed information on the requirements for all information elements in each file that composes a GTFS Schedule dataset (Figure 13).

3.1.2.3 GTFS Realtime

GTFS Realtime is a feed specification that allows public transport agencies to provide up-to-date information about current arrival and departure times, service alerts, and vehicle position, allowing users to smoothly plan their trips. Structure of the GTFS Realtime is shown in Figure 14.

[3] https://gtfs.org/documentation/schedule/reference/ [23.08.2024]

The specification currently supports the following types of information:

- Trip updates - delays, cancellations, changed routes
- Service alerts - stop moved, unforeseen events affecting a station, route or the entire network
- Vehicle positions - information about the vehicles including location and congestion level

GTFS Realtime was designed around ease of implementation, good GTFS interoperability and a focus on passenger information. This was possible through a partnership of the initial Live Transit Updates partner agencies, several transit developers and Google. The specification is published under the Apache 2.0 License.

Figure 14 GTFS Realtime structure.

The GTFS Realtime data exchange format is based on Protocol Buffers which is a language- and platform-neutral mechanism for serializing structured data (think XML, but smaller, faster, and simpler).

Similarly to GTFS Schedule, the GTFS Realtime Reference is the source of truth that establishes the rules and requirements for any GTFS Realtime feed, while the gtfs-realtime.proto file (Figure 15) defines the hierarchy of elements and their type definitions that are used.

```
// Copyright 2015 The GTFS Specifications Authors.
//
// Licensed under the Apache License, Version 2.0 (the "License");
// you may not use this file except in compliance with the License.
// You may obtain a copy of the License at
//
//     http://www.apache.org/licenses/LICENSE-2.0
//
// Unless required by applicable law or agreed to in writing, software
// distributed under the License is distributed on an "AS IS" BASIS,
// WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
// See the License for the specific language governing permissions and
// limitations under the License.

// Protocol definition file for GTFS Realtime.
//
// GTFS Realtime lets transit agencies provide consumers with realtime
// information about disruptions to their service (stations closed, lines not
// operating, important delays etc), location of their vehicles and expected
// arrival times.
//
// This protocol is published at:
// https://github.com/google/transit/tree/master/gtfs-realtime

syntax = "proto2";
option java_package = "com.google.transit.realtime";
package transit_realtime;

// The contents of a feed message.
// A feed is a continuous stream of feed messages. Each message in the stream is
// obtained as a response to an appropriate HTTP GET request.
// A realtime feed is always defined with relation to an existing GTFS feed.
// All the entity ids are resolved with respect to the GTFS feed.
// Note that "required" and "optional" as stated in this file refer to Protocol
// Buffer cardinality, not semantic cardinality.  See reference.md at
// https://github.com/google/transit/tree/master/gtfs-realtime for field
// semantic cardinality.
message FeedMessage {
  // Metadata about this feed and feed message.
  required FeedHeader header = 1;
```

Figure 15 GTFS Realtime -> gtfs-realtime.proto file

3.1.3 Data from mobile network (Telecom data)

Data collected from mobile network operators contains information on population migration and retention patterns. One can think about this data as digital breadcrumbs that can be patched up to present with low spatial accuracy users' locations. As the collected data sample is very large, analysis produced on it faithfully represent the migration patterns of the whole population.

Multiple mobile network data sources are collected and synthesised: Cell Trace Records (CTR), subscription database, and Performance/Configuration Management (PM/CM). Access to this data is given directly by the mobile operators and depends on their market share. All information is anonymised and scrutinized under privacy laws and GDPR compliance. Additionally, collection of this data is permitted for a limited time of one week, temporally separated by at least a month. Acquisition of this big data set is a time-consuming and sensitive process. From the set, one can approximate the position of mobile network users in time only,

additional analytical techniques are used to reconstruct the population movement (population trips) and general migration patterns. Based on the characteristics of these movements (time of occurrence, approximate distance, speed, duration, etc.), additional context such as purpose and mode of travel can be assigned.

Unfiltered fully anonymized mobile network data is ingested through an SFTP protocol and processed in a containerized Apache Spark/Hadoop environment. A schema of the process from data ingestion to data processing and finally, reporting or visualization is shown in Figure 16. Interaction with the Spark layer is conducted by a PySpark Python API, both due to ease of use, and ease of integration. Processed data is stored as parquets in an easily accessible Hadoop Distributed File System (HDFS). Before processing, due to the varying complexity, runtime parameters need to be finely tuned to optimally allocate cluster resources. This is needed because each of the big data sets of the collected mobile network snapshot can have different requirements that depend on the use cases. The computing infrastructure consists of a data cluster which runs 6 virtual machines. It has an 8 core CPU with 32GB of RAM and 1TB disk space. Primary software used is Spark, Python and Jupyter. One week of the collected mobile data can generate more than 6TB of information. The whole processing time to extract migration patterns and generate origin-destination matrices (ODM) can take a couple of days.

In Figure 16, the “Spark” step contains different scripts that refine and improve the spatial accuracy. First, data is prepared and filtered. Any outliers are detected and removed. Additionally, smoothing and averaging is conducted, and finally, validation and verification from outside sources is implemented. These sources are usually testing devices that log GPS data for better accuracy. The whole process is depicted in a schema shown in Figure 17.

Once locations are determined, positional data is further processed with complex algorithms to identify the type of user locations. Based on the logging activity of the mobile user and its timestamp we can discern between periods when the user is active, *i.e.*, moving or when his location does not change and is inactive. This information is flagged as *moving* or *stationary*, respectfully. If the activity is *stationary* further deduction can be made based on the clustering of location and time data for a certain user. These criteria, as shown in Figure 18, can give information on home, work, or other locations of users.

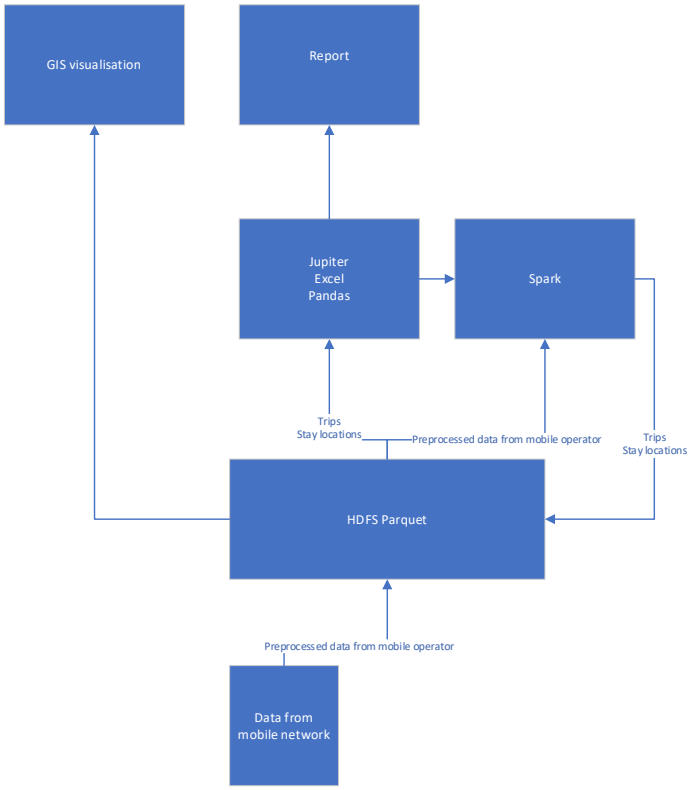

Figure 16 Mobile data ingestion and processing schema.

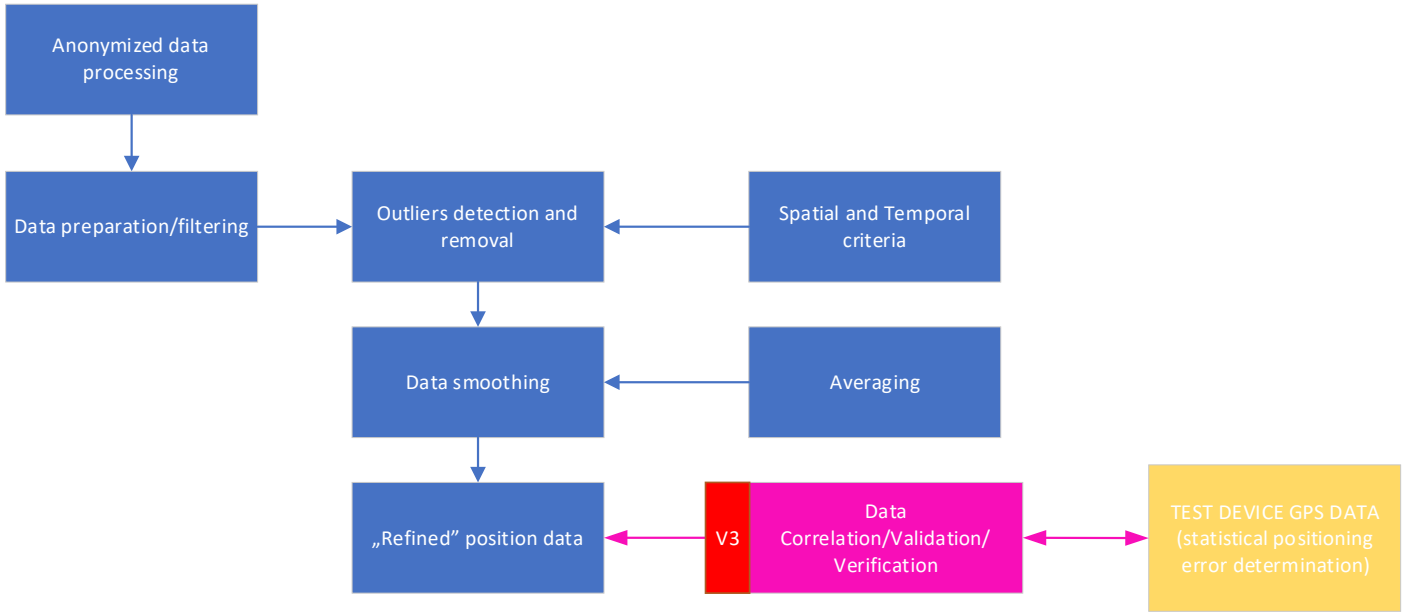

Figure 17 Processing steps for anonymized mobile data.

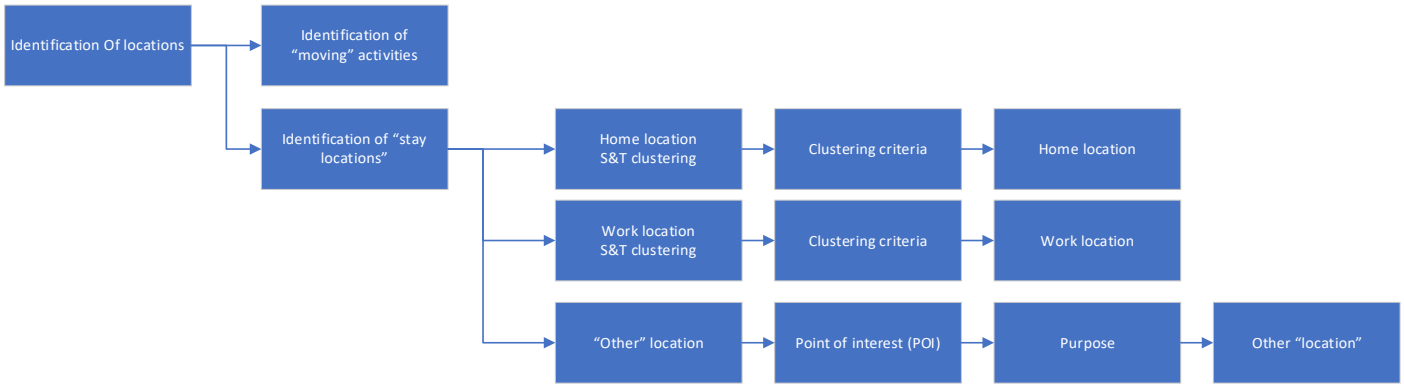

Figure 18 Processing schema for user location type deduction.

Figure 19 Processing steps for defining trip objects and identification of user migrations.

For use cases in this project, the territory of Croatia was segmented into sectors. These sectors were constructed based on information gathered by stakeholders on migration patterns and local infrastructure, and with regards to the local mobile network infrastructure. Once data is processed and sectors are assigned to the locations of mobile network users, objects like *staylocations* and *trips* are constructed. These objects contain information on sectors that were visited by users at a certain time, and if their location was stationary for more than a designated time (usually 15 minutes). In general, we define trip related parameters and criteria based on the preset temporal, spatial and speed thresholds. Once these thresholds are met *trip start*, *end* and *duration* can be identified. A detailed diagram of processes involved in defining *trip* parameters is shown in Figure 19. Further deductions can be made following the characteristics and patterns of a trip on the trip purpose, distance, and mode. Hence, *trips* data in the context of the previously mentioned use cases describe mobile users travelling to or from the Zagreb airport during the time of network recording. From *trips* we build origin-destination matrices (ODM) which allow us to further visualize and analyse migration patterns.

Mobile data also allows us to discern the country of origin of users. For the use case of airline strategic planning this is valuable information. It helps understand different modes of travel used by locals and tourists. This could help better public transport option to and from the airport. It also gives statistics on passengers' country of origin which can improve areas like flight connection planning and tourism advertising.

3.1.3.1 Identification of user migrations and user types

Anonymized telecom data is utilized to identify all users *staylocations*, which are then converted into user migrations (*trips*). The *staylocations* table contains, for each anonymized user, information on the duration spent in a specific geographical area. A universally unique identifier (UUID) is generated to connect items in the *staylocations* table with items in the *trips* table from which it is derived. User *trips* contain information on their starting and ending locations, as well as the start and end times of each trip, and information about the distance between the trip's starting and ending points.

Data used for analytics is, as mentioned, stored in a parquet format in a HDFS. This file format is language independent and has a binary representation. Parquet is generally used to efficiently store large data sets. It is a column-oriented file format, meaning that the data is stored per column instead of only per row. The parquet files are structured and include the schema of the columns, so it is suited for importing straight into a database/data warehouse. The general schema of the *staylocations* (Table 4) and *trips* (Table 5) data is given in the following tables.

Table 4 Data schema of the staylocations parquet.

Column name	Data type	Description
imsi	string	Encrypted International mobile subscriber identity (IMSI) value
start_time	timestamp	Timestamp of the beginning time when a user location was stationary
end_time	timestamp	Timestamp of the ending time when a user location stopped being stationary
last_known_time	timestamp	Timestamp of the last known time a user was registered to the network during the time his location was stationary
stayduration	integer	Duration of the time during which a user was stationary expressed in seconds

type	string	Can be *stationary* or *moving* which depends on the preset time taken when defining a staylocation (eg. User didn't change his location for more then 15min)
lat	float	Calculated latitude value of the user location
lon	float	Calculated longitude value of the user location
sector_id	integer	Sector ID in which the user coordinates fall into
country	string	Name of the country acquired from a country code the network detects when a user device registers

Table 5 Data schema of the trips parquet.

Column name	Data type	Description
imsi	string	Encrypted imsi value
trip_start	timestamp	Timestamp of the beginning of the defined trip
trip_stop	timestamp	Timestamp of the ending of the defined trip
trip_duration	integer	Total time of the trip expressed in seconds
start_lat	float	Latitude value of the location where the trip began
start_lon	float	Longitude value of the location where the trip began
end_lat	float	Latitude value of the location where the trip ended
end_lon	float	Longitude value of the location where the trip ended
start_sector_id	integer	ID value of the sector where the tip began
end_sector_id	integer	ID value of the sector where the trip ended
speed	double	Evaluated average speed during the trip in m/s for Euclidian distance
distance	integer	Air distance crossed during the trip in meters

visited_locations	array	Array of structures which contain information on the latitude, longitude, sector_id and timestamp of the locations visited during the trip

Using this data, the initial step involves identifying different user types. Methodology is depicted in Figure 20. The following categories are established: airport employee, airport user—passenger—arriving by aircraft, airport user—passenger—departing by aircraft, airport user (greeter/buyer), airport user taxi driver, individuals passing near the airport—local residents—and commuters. Each of these user types can be identified based on their characteristic behaviour, using criteria such as the identification of their Home/Work/Other sector, the duration of their stay at the airport sector, trip characteristics, user type (domestic or foreign), and the time of day, among others.

For instance, a potential candidate for an airport employee could be someone whose estimated home sector is not the airport but spends a significant amount of time in the airport sector during their day. An airport employee is expected to have a pair of trips originating from and departing to the airport, and these trips should occur during airport working hours. Conversely, a taxi driver is anticipated to have a substantial number of trips originating from and departing to the airport, with the retention time at the airport not being significant. In the case of a passenger arriving at the airport, their first occurrence should be at the airport sector, while a departing passenger should have their last occurrence at the airport sector. Following the identification process, all these user types can be verified using various data sources, as depicted in Figure 20. Subsequently, all trips in the trip database can be categorized under the appropriate user type, allowing for consideration of only the trips relevant to a specific user type for each use case.

As mentioned earlier, the transport mode can be determined based on trip characteristics such as origin, destination, duration, speed, and distance. For this study, a methodology has been implemented to distinguish between the two most common transport modes used to travel to the airport—personal vehicles and public transport. This proposed methodology relies on a set of rules and the application of Bayes statistical modelling, as outlined in [23]. Subsequently, all trips associated with identified user types are then categorized with the appropriate transport mode. The accuracy of the transport mode data can be validated using supplemental datasets from the airport parking and data from traffic counters in the airport area's road network.

Upon completion of this process, verified migration patterns have been produced and stored in a suitable database, thus enabling the fulfilment of the previously identified use cases. Validation data sources have been recommended and regarded as the ground truth data for each data source and every measured or identified parameter. Following the analytical process, the analytical findings will be validated using the ground truth data to pinpoint any inaccuracies in the results.

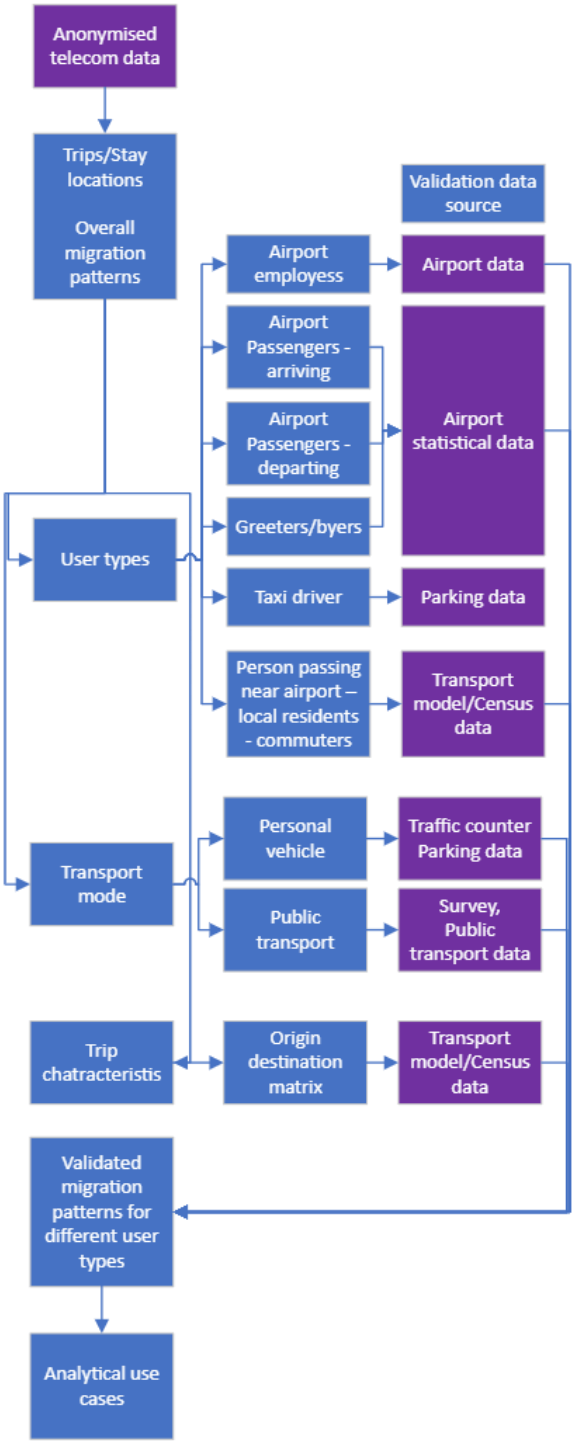

Figure 20 Methodology

3.1.4 Airport data

3.1.4.1 Flights history data for Zagreb airport

Flights history data is weekly retrieved from a shared OneDrive folder from Zagreb airport. Every Monday data from the previous week is uploaded. Description of data can be seen in the following table.

Flight history data loaded from Zagreb airport's OneDrive folder	
Date	Flight date
STA	Scheduled time of arrival
STD	Scheduled time of departure
Dir.	Direction: arrival or departure
Country sto. 1	Country of destination 1
Block dt	The aircraft shuts down its engines after landing or when it starts its engines before take-off (date)
Block	The aircraft shuts down its engines after landing or when it starts its engines before take-off (time)
Tr.T dt	Runway time – the date when the aircraft touched the runway when landing or when it separates from the runway during take-off
Tr.T	Runway time – the time when the aircraft touched the runway when landing or when it separates from the runway during take-off
Nat.	Flight nationality: domestic, European – international
Transfer Infants	Number of transfer infants
Pay.Pax.1	Number of passengers per stopover or destination 1
Pay.Pax.2	Number of passengers per stopover or destination 2

PAD 1	Number of transfer passengers per destination 1
PAD 2	Number of transfer passengers per destination 2

3.1.4.2 Flights history data for Charles de Gaulle airport

Flights history data cover a time span from 1st February until 26th March 2024, while parking data was gathered from 1st January until 31st March 2024.

Because of unequal timeframes, the used dataset contains data collected from 1st February until 26th March 2024 since both sources are needed for further analysis.

Automatic updating of data from Charles de Gaulle airport was not possible. However, it is possible to use the prediction algorithm for the number of vehicles entering or exiting CDG's parking lots. It would be possible to setup another data transfer pipeline so the data for previous week could be sent to our data management platform automatically on the 1st day in the week. After that the ML model retraining would be started and prediction data for the current week could be generated.

Flights history data are not aggregated so we must aggregate them in the first step, while parking data we received from CDG airport have already been aggregated on hourly basis.

Flights history data is described in the following table (Table 6).

Table 6 Schema of the flight history data from CDG airport.

Flight history data from Charles de Gaulle airport	
Type de mouvement (Type of movement)	Direction of flight: arrival or departure
Jour (Day)	Date
Horaire théorique (Theoretical schedule)	Scheduled date and time of arrival or departure
Horaire bloc (Block schedule)	The aircraft shuts down its engines after landing or when it starts its engines before take-off (date and time)

Temps (piste). Horaire piste (Track time. Track schedule)	Runway time – the date and time when the aircraft touched the runway when landing or when it separates from the runway during take-off
Ville (City)	City of airport from/to which an airplane is flying
Ville_ref (City_ref)	City of airport from/to which an airplane is flying
Pays (Country)	Country of airport from/to which an airplane is flying
Nombre de passagers réalisé (Number of passengers carried out)	Total number of passengers carried out
Nombre de passagers réalisé en correspondance (Number of connecting passengers)	Number of passengers who are connecting to another flight (CDG airport is not their final destination).

3.1.5 Parking data

3.1.5.1 Parking data for Zagreb airport

Parking data is updated daily and can be retrieved from a shared OneDrive folder from Zagreb airport. Details on the collected data shown in the following table (

Table 7).

Table 7 Parking data schema.

Parking data loaded from Zagreb airport 's OneDrive folder	
Occurred	The time of passing through the ramp

EventDescription	Description of the event. It can have one of two values: - Ulazak korisnika (User entry) - Izlazak korisnika (User exit)
Location	Designation consisting of multiple numbers, which can be translated as the number of device (the ramp) followed with slash and unique number that describes one user (e.g. 21/28270281123023400057971)
DeviceName	The name of the parking lot ramp
Note	Additional note if it exists

As it was already mentioned, because of unequal timeframes, we only use the part of the data for the time span from 1st February until 26th March 2024. This is because we need data from both sources (flights history and parking data) with matching timestamps.

To be able to analyse flights history and parking data, the data is aggregated on an hourly basis separately. After that, aggregated data is combined and finally stored in the database. Data that is ready to be saved into the data management platform is described in the following table (Table 8).

Table 8 Aggregated and combined flight history and parking data.

Hourly aggregated flights history and parking data stored into database	
Vrijeme (Time)	The time of passing through the ramp
<parking_lot_name>_total_IN	Total number of vehicles entering parking lot within one hour. <parking_lot_name> can be any of the following parking lots: B2B, GP, RENT-A-CAR, TAXI, VIADUKT
<parking_lot_name>_total_OUT	Total number of vehicles exiting parking lot within one hour.
<parking_lot_name>_num_of_vehicles	Total number of vehicles in the parking lot at the end of the specific hour interval. Number of vehicles entering parking lot in the current hour

	is added to the number of vehicles in the parking lot for the previous hour and after that the number of vehicles exiting the parking lot is subtracted from the previously calculated sum.
Parking_total_IN	Total number of vehicles entering all parking lots within one hour.
Parking_total_OUT	Total number of vehicles exiting all parking lots within one hour.
PAX1_ARR	Number of passengers for stopover or destination 1 arriving to Zagreb airport within one hour.
PAX2_ARR	Number of passengers for stopover or destination 2 arriving to Zagreb airport within one hour.
PAD1_ARR	Number of transfer passengers for destination 1 arriving to Zagreb airport within one hour.
PAD2_ARR	Number of transfer passengers for destination 2 arriving to Zagreb airport within one hour.
PAX1_DEP	Number of passengers for stopover or destination 1 departing from Zagreb airport within one hour.
PAX2_DEP	Number of passengers for stopover or destination 2 departing from Zagreb airport within one hour.
PAD1_DEP	Number of transfer passengers for destination 1 departing from Zagreb airport within one hour.

PAD2_DEP	Number of transfer passengers for destination 2 departing from Zagreb airport within one hour.
traffic_participants_ARR	Total number of traffic participants for arrival flights within one hour.
total_pass_ARR	Total number of passengers arriving at the Zagreb airport within one hour.
traffic_participants_DEP	Total number of traffic participants for departing flights withing one hour.
total_pass_DEP	Total number of passengers departing from the Zagreb airport within one hour.

After that, model for prediction of parking lot usage is being retrained and prediction data for current week is being generated.

All of this starts by running a Python script on an Oracle cloud instance which carries the data management platform. After the script for fetching data is successfully ran and new data is retrieved, the next Python script must be started. This script aggregates on an hourly basis the received raw data for flights history and parking, it is then merged into one data set and stored into the database. This data can then be retrieved from the platform via REST API. Current configuration is set to only allow the administrator to have permissions to fetch this data via REST API. If needed, other users can also be given required permission to access aggregated data for parking and flights history.

Files with raw data are stored in the separate folder that is not accessible by any other user except the Oracle Cloud instance's administrator. This data is deleted after 6-months period.

3.1.5.2 Parking data for Charles de Gaulle airport

Parking data from Charles de Gaulle airport were received in a form of Excel file. This data will not have a continuous delivery or update with new data. Data from CDG is aggregated on an hourly basis, unlike data from Zagreb airport. Data covers the time window of 3 months, starting at January and ending with March of 2024. Structure of the received data is described in the following table (Table 9).

Table 9 Schema of the parking data from CDG airport.

Parking data from Charles de Gaulle airport

Année (Year)	Year
Mois (Month)	Month of the year
Date	Date in format d.m.yyyy
Heure (Hour)	The hour mark for which the data is valid (e.g. 01h for data between 00:00h and 01:00h)
Famille de parcs (Family of parks)	Family/type of parks
Parking	Parking lot name
Groupe de client (Customer group)	Customer group
Entrées (Entries)	Number of vehicles entering parking lot within one hour
Sorties (Exits)	Number of vehicles entering parking lot within one hour

To be able to use received flights history and parking data, flights history is also aggregated on an hourly basis separately. After that, the aggregated data is combined and finally stored in the database. Data ready to be saved into data management platform is described in the following table (Table 10).

Table 10 Shema for combined flights history and parking data aggregated on an hourly basis.

Hourly aggregated flights history and parking data stored into database	
DateTime	Date and hour of day for which the data is valid.
<parking_lot_name>_total_IN	Total number of vehicles entering parking lot within one hour.
<parking_lot_name>_total_OUT	Total number of vehicles exiting parking lot within one hour.

Parking_total_IN	Total number of vehicles entering all parking lots within one hour.
Parking_total_OUT	Total number of vehicles exiting all parking lots within one hour.
PAX1_ARR	Number of passengers for stopover or destination 1 arriving to Charles de Gaulle airport within one hour.
PAD1_ARR	Number of transfer passengers for destination 1 arriving to Charles de Gaulle airport within one hour.
PAX1_DEP	Number of passengers for stopover or destination 1 departing from Charles de Gaulle airport within one hour.
PAD1_DEP	Number of transfer passengers for destination 1 departing from Charles de Gaulle airport within one hour.
traffic_participants_ARR	Total number of traffic participants for arrival flights within one hour.
total_pass_ARR	Total number of passengers arriving at the Charles de Gaulle airport within one hour.
traffic_participants_DEP	Total number of traffic participants for departing flights withing one hour.
total_pass_DEP	Total number of passengers departing from the Charles de Gaulle airport within one hour.

<parking_lot_name> in the previous table can be any of the following parking lots: CDG-ADM, CDG-BDM, CDG-CDM, CDG-DDM, CDG-EDM, CDG-EMPORT-E, CDG-FDM, CDG-LIN1, CDG-LINE, CDG-LINF, CDG-P1, CDG-P1DM, CDG-P3, CDG-P3DM, CDG-P3-RESA, CDG-PAB, CDG-PCD, CDG-PEF, CDG-PG, CDG-PGA, CDG-PGDM, CDG-PH, CDG-PJ, CDG-P-P1, CDG-P-PAB, CDG-P-PCD, CDG-P-PEF, CDG-P-PG, CDG-PR, CDG-PROABVL, CDG-PROCD, CDG-PROCDG1HG, CDG-PROCDG1VL, CDG-PROEFVL, CDG-PROFHG, CDG-PW, CDG-PX, CDG-PZ.

After the data is stored into the data management platform, model for the prediction of parking lot usage is retrained and the predicted data for the current week is generated.

This process is the same as for Zagreb airport data. Only difference can be found in setting up the parameters for a sent request to the data management platform that starts the Apache Spark job.

When the Spark job generates predictions for every parking lot, it stores the results into data management platform.

This data can be retrieved via REST API in Simplified JSON, JSON and XML formats. Table 11 shows a comparison between Simplified JSON and standardized JSON formats. For every data (value) there's additional data about source to which this data belongs. Except address part, there's also information about data type that is stored as a value and timestamp. It's important to note that the time is in UTC time and the format used in output is yyyy-MM-ddTHH:mm:ssZ where yyyy describes a year, MM describes ordinal number of a month in a year written in two-digit format, dd describes ordinal number of a day in a month written in two-digit format, letter T separates date and time numbers, HH describes a hour in a day in range from 00 to 24, mm describes a minute in two-digit format (00-59), ss describes a second number while letter Z describes time zone offset of 0 hours (no offset from UTC time).

Table 11 Example of forecast data of incoming traffic to parking lot GP on Zagreb airport (value pair MZLZ, i.e. "Međunarodna zračna luka Zagreb") retrieved via REST API

Simplified JSON	**JSON**
{ "cn": [{ "so": { "dggs": "MZLZ", "dgws": "ParkingData", "sens": "ParkingForecast", "ress": "GP_total_IN_forecast" }, **"v": 42,** **"ty": "integer",** **"t": "2023-09-07T21:00:00Z"**	{ "contentNodes": [{ "source": { "gatewayGroupSpec": "MZLZ", "gatewaySpec": "ParkingData", "sensorSpec": "ParkingForecast", "resourceSpec": "GP_total_IN_forecast" }, "value": 42, "type": "integer", "time": "2023-09-07T21:00:00Z"

<table>
<tr>
<td>

<pre>
    },
    {
      "so": {
        "dggs": "MZLZ",
        "dgws": "ParkingData",
        "sens": "ParkingForecast",
        "ress": "GP_total_IN_forecast"
      },
      "v": 47,
      "ty": "integer",
      "t": "2023-09-07T20:00:00Z"
    },
    ...
  ]
}
</pre>

</td>
<td>

<pre>
    },
    {
      "source": {
        "gatewayGroupSpec": "MZLZ",
        "gatewaySpec": "ParkingData",
        "sensorSpec": "ParkingForecast",
        "resourceSpec": "GP_total_IN_forecast"
      },
      "value": 47,
      "type": "integer",
      "time": "2023-09-07T20:00:00Z"
    },
  ...
  ]
}
</pre>

</td>
</tr>
</table>

cURL requests for data retrieval from Table 11 are described in the Table 12.

Table 12 Example of cURL requests for retrieval of forecast data

Retrieval of data in Simplified JSON format	**Retrieval of data in JSON format**
curl -X GET -u <username>:<password> http://localhost:8181/m2m/data?resourceSpec=GP_total_IN_forecast -H "Accept: application/vnd.ericsson.m2m.output.short+json"	curl -X GET -u <username>:<password> http://localhost:8181/m2m/data?resourceSpec=GP_total_IN_forecast -H "Accept: application/vnd.ericsson.m2m.output.long+json"

It's visible that there is just a small difference between two requests: “Accept” header in the request. This header defines which output adapter in Data Management Platform will be used for generating output.

3.1.6 Other data

A questionnaire for the Zagreb airport employees was carried out. Results are available on a shared OneDrive folder from Zagreb airport. An example of questions asked in the questioner is shown in Table 13. Questions posed were about the employee's habits and which means of transport they usually use to get to work. The questionnaire was anonymous, but gathered information on employees such as the city where the employee lives, how they come to work (car/ bus/ bicycle) and preferred means of travel to work.

Table 13 Example of question on the questioneer.

Questionnaire for employees (only questions about other ways of transport)	
Ukoliko se obično vozite sami na posao, koju biste od sljedećih alternativa za putovanje na posao uzeli u obzir barem jedan dan u tjednu? Označite sve što se odnosi.	If you usually drive yourself to work, which of the following commuting alternatives would you consider at least one day a week? Check all that apply.
Ukoliko se obično vozite sami na posao, koji od sljedećih poticaja bi vas potaknuo da koristite pogodnosi putovanja alternativnim načinom barem jedan dan u tjednu. Označite sve što se odnosi.	If you usually drive yourself to work, which of the following incentives would encourage you to use the convenience of alternative travel at least one day a week. Check all that apply.

Another questionnaire was conducted on users that come by car. Questions were similar covered their interest in using other means of transport to get to the airport.

Faculty of Transport and Traffic Sciences conducted monitoring and counting of the entry and exit of people using the available city transport on line 290. Data contains the number of people that had entered and exited the bus on line 290 (Kvaternik square – Velika Gorica). One bus stop along this line is in front of Zagreb airport and data for that stop were used for the analysis. This covers a time period between 28th of September 2022 to 30th of September 2022.

Google maps was used to estimate the arrival time from the airport to the Zagreb city centre.

For Paris aerodrome Charles de Gaulle (CGD) information on public transport data was gathered from the web page Paris-Charles de Gaulle airport by public transport - Paris Aéroport (parisaeroport.fr).

We also used information on the national holidays from the web page https://mojkalendar.com.hr/2022.aspx which was useful for data validation. For parking analysis and forecast of traffic flow through parking lots it's important to know distinguish between a working day, a weekend dan and a holiday. List of holidays in France is retrieved from web page https://publicholidays.fr/2024-dates/.

All mentioned data is analysed with help of Python programming language, Pandas and Matplotlib libraries and is written in a form of Jupyter Notebooks. A ZIP file consisting of Jupyter Notebooks with additional data is created and saved into Data Management Platform so that users with access to the Platform can easily retrieve it and review all the results.

3.2 Data ingestion mechanisms

There are several input and output adapters which support various data formats. Probably the best-known formats for data exchange are XML and JSON, so our platform supports data ingestion and retrieval using both formats. There are several standards which use their own data formats. Our platform supports the following:

- DATEX II: The information model for exchanging road traffic and travel information in Europe.[4]
- NeTEx: CEN technical standard for exchanging public transport schedules and related data.[5]
- GTFS (General Transit Feed Specification): Community-driven open standard for rider-facing transit information. It was started with collaboration between TriMet and Google. They worked together to format Google's transit data into an easily maintainable and consumable format that could be imported into Google Maps. This transit data format was originally known as the Google Transit Feed Specification.[6]
- SIRI: Service interface for real-time information related to public transport operations. Like NeTEx, SIRI is also a CEN technical standard. It provides an abstract model of common public transport concepts and data structures that enables the exchange of information on transport operations between different computer systems. SIRI was established as European standard in October 2006.[7]

Most Aggregated and combined flights history and parking data are stored (ingested) into the data management platform via REST API.

[4] https://datex2.eu
[5] https://netex-cen.eu
[6] https://gtfs.org/about/
[7] https://www.siri-cen.eu

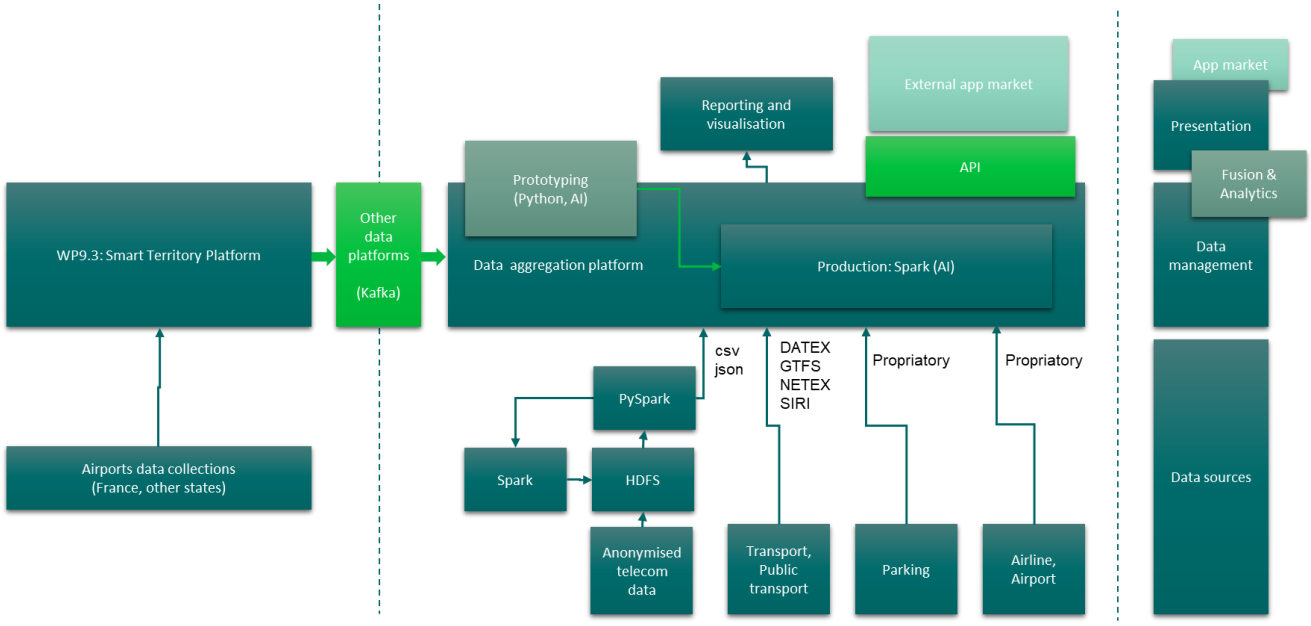

Figure 21 System Architecture.

System architecture shown in Figure 21 describes the main parts of the data aggregation platform and the sources from which data is aggregated and outputs generated by the platform.

Anonymised telecom data has the biggest footprint. Its size can be in an order of terabytes. One week of telecom data collection reaches a size of around 6 TB. Raw data is collected by different probes mounted on the telecom operator hardware and written on hard disks. It is further processed and correlated before the disk with the data can be mounted to our cluster. Once mounted it is stored in a parquet format within a HDFS (Hadoop Distributed File System) instance. Python scripts relying on the PySpark library (Python API for Spark) guide the Apache Spark cluster for further large-scale data processing. Specific databases stored as parquets in the HDFS are generated by these scripts which can then be used for further analysis. This analysis is done through Jupyter Notebook using the Python programming language. The result of analysis is generating needed information which are able to map migration patterns of mobile network users (origin-destination matrices). Data preparation for the analysis can take about 48 hours on a cluster consisting of 6 virtual machines, where each virtual machine has 8 CPU cores, 32 GB of RAM and 1 TB of disk space.

For the requirements of this work package the results of the analysis were prepared as the following two files:

- JSON file with recorded trips of users that visited Zagreb airport during the week of recording telecom data. These trips cover taxi drivers servicing the airport and passing through the airport's parking lot, employees working at the airport, passengers on departing flights and passengers on arriving flights.
- CSV file containing information on the nationality of every user. This is deduced from the International mobile subscriber identity (IMSI) which contains the mobile country code (MCC).

Additional libraries utilized with Jupyter Notebook for data analysis are Pandas and Matplotlib. Available functionalities in these packages were used to generate graphs and maps that shows passengers migrations. The aim was to visualize (migration patterns) trips taken by passengers before arriving to the airport for their departing flights and trips from the airport after passengers arrived on a flight.

Public transport data was collected in DATEX II, NeTEx, GTFS and SIRI formats and further aggregated.

Parking data from Zagreb airport is continuously collected in a proprietary format (CSV file), while parking data from Charles de Gaulle airport was prepared and uploaded as an Excel file.

Flights history data contains the flights schedule and number of passengers on each flight. It is collected in a proprietary format (Excel file), both from Zagreb and Charles de Gaulle airport.

The parking and flights history data received from Zagreb airport is set up to automatic exchange and refresh data via OneDrive.

Flights history and parking data are aggregated on an hourly basis and correlates using a Python script. After this processing it is stored into the Data Management Platform in a searchable format. Apache spark jobs are deployed on a weekly basis after data is updated, containing new data from the previous week. The Spark job trains the ML model on historical hourly data containing the number of vehicles passing through the parking lots and generates forecasted data of the same type for the current week.

Each approved user of the Data Management Platform can retrieve the forecasted data in a JSON format through a REST API shown as a light green box on the Figure 21. The setup is suitable for M2M (machine to machine) connections so the data can be requested and used by other applications with approval to retrieve data from the Data Management Platform.

Analysis of the telecom data and its connection with the airport data on arriving/departing flights was conducted within Jupyter Notebooks using python scripts. Graphs generated through the analysis and all scripts within the Jupyter Notebooks were compressed into a ZIP format and uploaded into the Data Management Platform. These files are available to anyone with a valid account and proper permissions via REST API.

Restrictions on accessing the forecasted parking data for Zagreb and Charles de Gaulle airports were put in. Specific rights are given to each user to view either the Zagreb airport data or the Charles de Gaulle airport data but not both.

The Data Management Platform enables simple integration and reception of data streams from other similar platforms via the Kafka interface, shown by the light green box in Figure 21. If needed an equally viable option would be using MQTT broker instead of Kafka. Within OLGA WP9.3, a Smart Territory Platform contains various data sets that could be useful for multimodal analytical use cases, and it could be integrated in the mentioned way in the future.

4 Analytical use cases development and results

4.1 Zagreb Airport

In this chapter, capabilities of the newly developed “IT solution (platform) for traffic flow optimization” application will be demonstrated and presented using priorly mentioned data sets.

4.1.1 Public transport optimization for nearby residents and airport users (both passengers and employees).

The goal of this use case is to optimize existing public transport lines and/or introduce new public transport lines in the Zagreb airport catchment area which would serve residents, passengers, and airport employees. Overall benefit is the availability of “IT solution (platform) for traffic flow optimization”, as a tool for general public transport optimisation and a shift towards sustainable landside transportation. Capabilities and screenshots from the application are presented on Figure 22 and Figure 23.

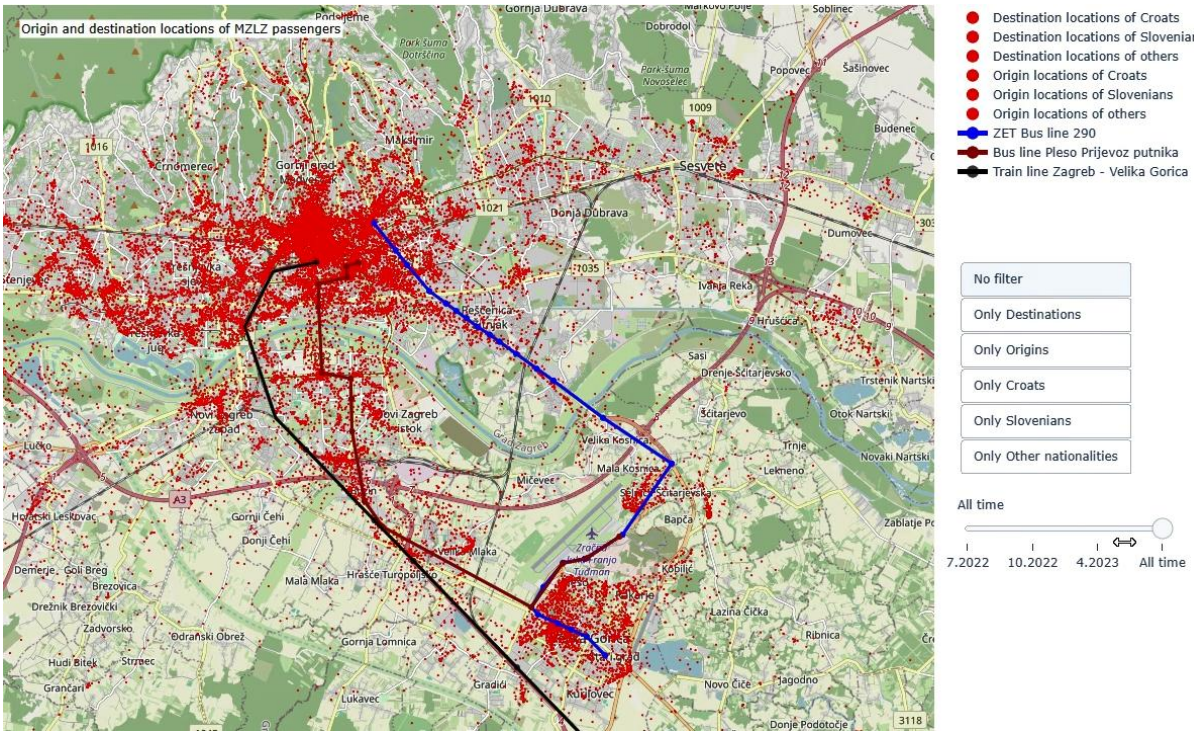

Figure 22 Origin and destination locations of Zagreb airport passengers. Lines mark three available public transport lines.

Along with the conventional data sets on public transport, telecom data had a crucial role in elaborating analytical use cases, because it enabled the identification of user types, that along with other data sources, helped give a better understanding of the traffic optimization problems of Zagreb airport.

Zagreb Airport is served by three lines of public transport, shown as coloured lines on Figure 22. Blue line represents the Zagreb Electric Tram (ZET) bus line 290, red line covers the Pleso bus rides and black line is the train line. Airport is not served by a tram line. Therefore, those three lines represent all scheduled public transport options for reaching the airport.

The results of the analysis are presented on Figure 22. Red dots on the picture mark locations of origins and destinations of trips starting or ending at the airport. This information was gathered from the trip objects generated from the telecom data. The analysis has shown that public transport stops near the airport are not optimised, and therefore, some public transport options are not considered by airport employees and users as appropriate option. Visualization shows that the train station is farther from the airport than the bus station. The nearest train station is the station "Velika Gorica" and from that station airport is only reachable by taxi. Possibility of introducing a train station near airport should be considered, since the main train station is at Zagreb city centre, where, according to telecom data, highest density of airport users occurs.

The location of the bus stop of line 290 is in front of Zagreb airport. Buses are convenient, and currently represents the best option to reach the airport by public transport. The location of the public transport lines and other stops, when compared with transport demand from telecom data, shows that the most frequent locations of origins and destinations for user trips starting or ending at the airport (transport demand) does not match with the current transport offer. This shows room for bus line adjustments. Additionally, traveling from the airport to the Zagreb centre takes around one hour by bus with multiple stops along the way. A direct line from the airport to the centre of Zagreb would reduce travel time. Also, repositioning of existing lines and optimisation of timetables should be considered to harmonise transport offer and transport demand. This visualisation is an example of how the optimisation tool can be used as a decision support tool for strategic planning and making transport policies. One of the conclusions made based on this analysis might be that there is a need for a new railway line that would go directly to the passenger's terminal on the airport. It also shows how the "Pleso" bus is much more frequently used when compared with the public bus (ZET bus line 290).

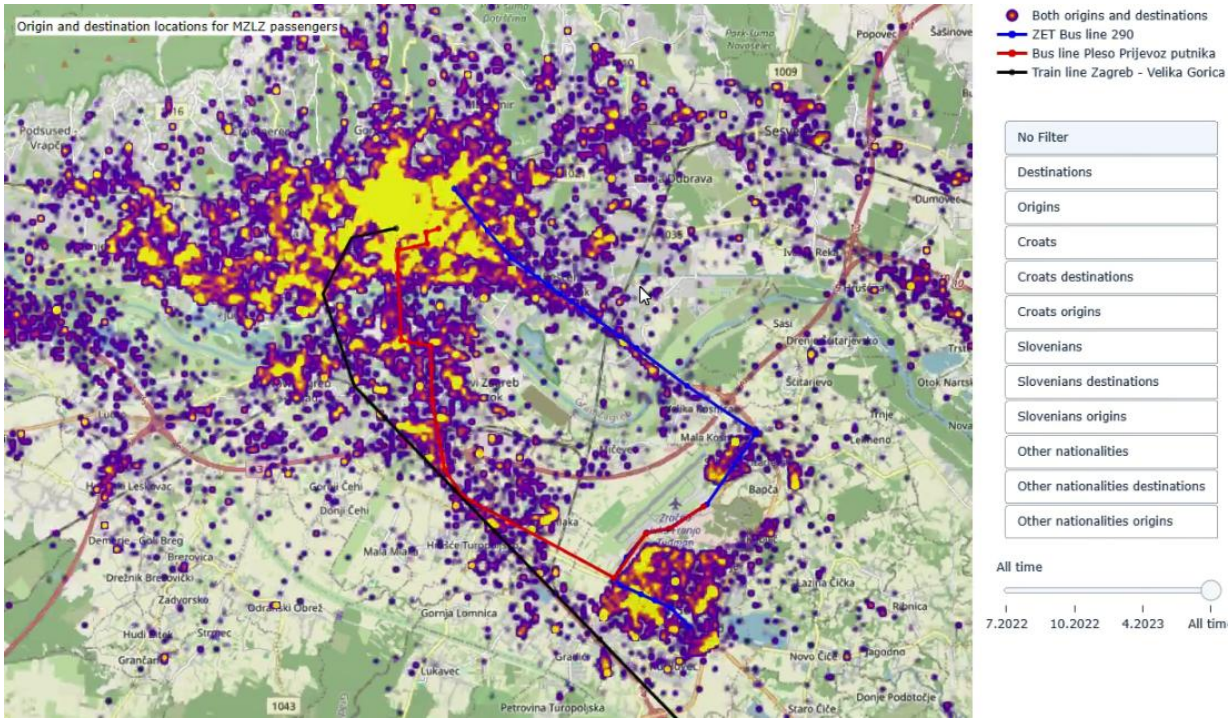

Figure 23 Heat map of origins and destinations.

This IT solution is a developed analytical tool that can perform various kind of analytics, including filtrations as required by a specific use case. The filter options are presented on the right side of the visualization (e.g. see Figure 23). For example, the tool is capable of filtering users by type or specific characteristics (like nationality). An example of how it can be used by filtering data (by origins, destinations, or their combination) shown on a heatmap is presented on Figure 23. These filters can be used for tailoring specific transport policies and targeting the specific types of users. For example, since the Zagreb airport is near Slovenia, it also serves several users from Slovenia. In this example, this tool can identify how such users arrived at the airport, and this can act as a base for further optimisations targeting a specific user type. In future research, interactive IT solution optimisation tool will be configured to enable most common preset analytical use cases.

4.1.2 Strategic planning of the airport gravitational areas and catchment zones.

With this use case we address the topic of how an airport can increase competitiveness by identifying gravitational and catchment areas, by analysing, and enhancing sustainable transport connections to these areas, and by detecting other areas which lack a public connection to the airport. From mobile network data

we gain an insight to airport users' migration patterns and behaviour. This is done by analysing trips that have their origins and/or destinations at the airport.

Following figures (Figure 24, Figure 25) show screenshots from the "IT solution (platform) for traffic flow optimization" which help analyse passenger migrations, determine and explore the real catchment airport area on a micro (airport narrower area) or macro level (entire territory of Republic of Croatia).

Visualisation of processed data in the newly developed IT solution (platform) for traffic flow optimization (Figure 24) in the Zagreb's surrounding area shows user migrations originating or ending at the Zagreb airport. The total number of passengers in each sector is displayed with circles of various diameters proportional to the number. This visualization clearly shows the main gravitational areas and catchment zones. The filters on the right side of the view offer a few possible choices and modifications to the visualization. A user can choose between "both origin and destination" of passengers to be shown on this map, "only destinations" or "only origins". Depending on the filter preference, the number of passengers that start or end their journey in a specific sector will change. This is a useful way of discerning between, and analysing different gravitational areas, since the tool can help visualize passenger demand for various time slots (during the time data was collected – July 2022, October 2022, April 2023). Platform user can choose an appropriate time slot and get insights into migration patterns for various characteristic times during the year (touristic season, off season period, extended weekend, regular season…) etc.

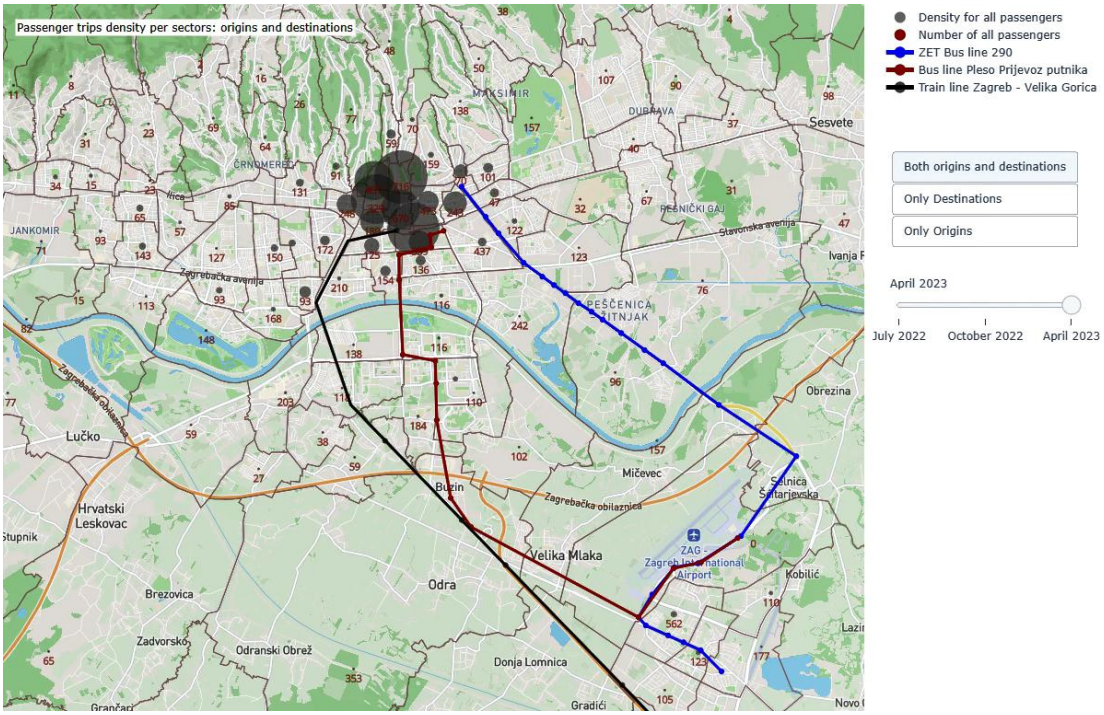

Figure 24 Passenger trips density per sector.

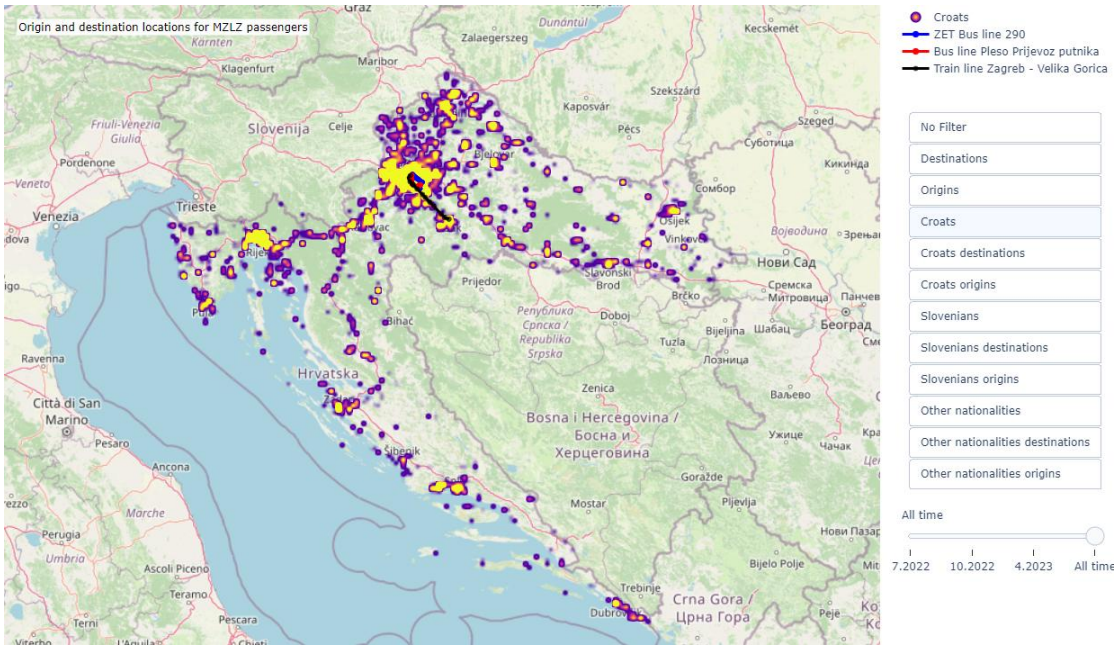

Figure 25 Heat map of origins and destinations in Croatia.

Our analysis of gravitational and catchment zones of Zagreb airport (origins and destinations of Zagreb airport users) is not limited to the Zagreb area. Telecom data was collected for all users in Croatia. Although the analysis has shown (Figure 25) that the major gravitational area and catchment zone of Zagreb Airport is in its surroundings, the heat map also shows that there are users travelling to and from Zagreb Airport from all points of the country (from cities like Rijeka, Split, Dubrovnik Osijek, Zadar…) even though those parts of Croatia have a closer airport. Therefore, we see how this analysis can provide a valuable insight for airport strategic planning in general. Since the telecom data analysis is limited to the coverage of the mobile network operator, we find that several trips originate near the Croatian border with neighbouring countries. This indicates that foreigners also use Zagreb airport for their travel needs, and that the Zagreb Airport catchment and gravitational areas extend outside of the Croatian territory.

In the future, additional telecom data, e.g. the call data records (CDR), can be added to the analysis to get information on domestic users outside of Croatia roaming on foreign mobile networks. This data can help understand why some users choose to use foreign airports (e.g. Ljubljana, Vienna, Budapest etc…) rather than the closer Zagreb airport in their home country.

4.1.3 Airline strategic planning

Analytical use case 3.3.3, Airline strategic planning, aims to give a better insight on airport travel demand and improve planning of airline operations. Analysis is carried out on information of the initial country of origin and destination of airline users. It requires both flights history data and telecom data. Flights history data gives us information on the country the departing flight was travelling to. However, telecom data contains information on the individual user's mobile country code (MCC). If the user was an airport passenger this code tells us about their home country, i.e. their nationality, which is the likely returning point for passengers on a departing flight.

We have combined information from both data sources and visualized it on an interactive map. The number of passengers and their home countries were aggregated daily for each day during the time mobile data was collected (a week in July and a week in October). Results of our analysis for both weeks are shown in two graphs, Figure 26 and Figure 27, respectively. The graphs show a map of Europe where each country is coloured in a blue or red tone. Countries from which a flight connection from Zagreb airport existed (served) on a certain day were marked with the blue colour. On the other hand, home countries of the departing passengers without a direct flight from Zagreb were coloured red. A darker tone of a certain colour indicates a higher number of departing passengers from that country.

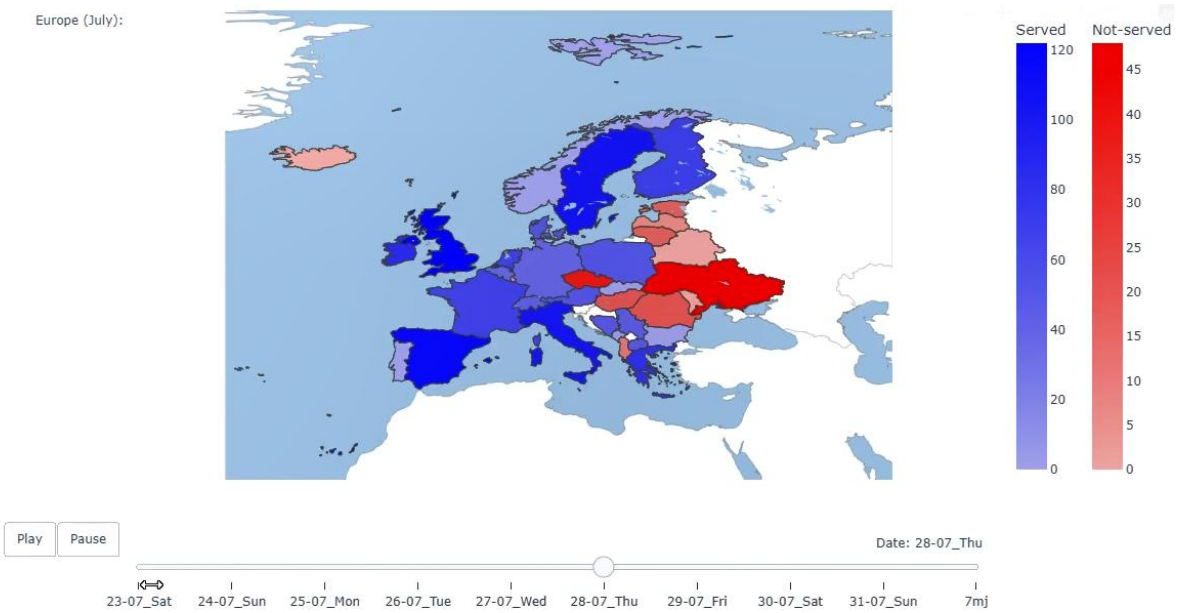

Figure 26 Served and not-served countries with a flight connection from Zagreb airport for specific day in July.

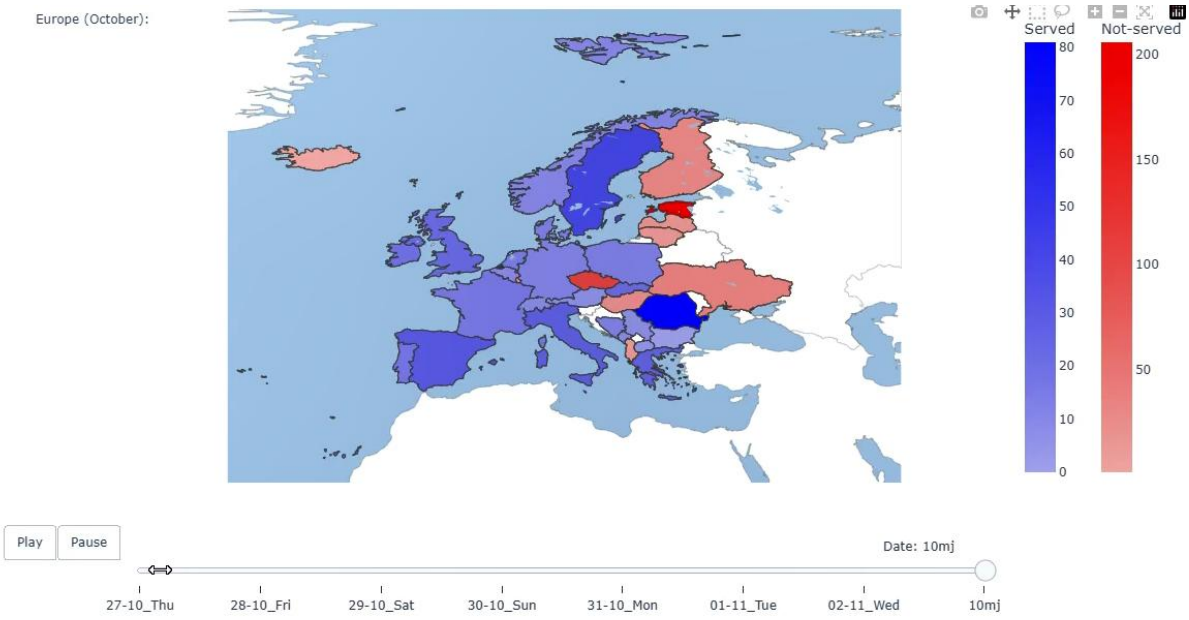

Figure 27 Served and not-served countries with a flight connection from Zagreb airport for specific day in October.

Those figures clearly indicate there is a significant travel demand from several countries (like Czech Republic, Estonia, some Eastern European countries, Island etc…) without a direct airplane connection, and that residents from those countries travel to Zagreb Airport via transfer flights or using other available options.

This analysis can provide valuable insights from which various stakeholders in the airport business chain can benefit from. Besides the Airport itself, this information is useful for airline companies, because it helps them identify and quantify true user demand for direct flights to and from Zagreb. Based on such information, airports can calculate a business case for modification and betterment of their transport offer, by modifying existing lines (introducing new or abandon old ones…).

Besides further refinement of the methodology, following steps might include extending the analytics by an inauguration of a new type of telecom data (CDR data) that will expand the analysis of domestic users which have origin/destination of the trip in Zagreb airport, but the final origin/destination in some other international airport. This would help in better identifying airports with a significant travel demand, but not served with a direct flight connection to Zagreb Airport.

4.1.4 Analytics of migration and retention habits of passengers.

Mobile network data from all three snapshots (of a week in July 2022, a week in October 2022 and a week in April 2023) was used to analyse retention habits of passengers. We have detected passengers travelling on departing flights from Zagreb airport and analysed their behaviour and retention time at the airport before the flight. A distinction was made between passengers on domestic flights and international flights based on a premise that passengers arrive earlier for international flights. Additionally, a hypothesis was proposed that the type of passenger could be distinguished based on their retention time at the airport. An assumption would be that users which fly on business use benefits of a faster check-in line and have priority boarding. They would likely arrive later for their flight and spend less time at the airport.

Histograms showing the total time passengers spent at the airport before their flight were generated for all available data. Two plots were made for each available data period, each for passengers departing on domestic and on international flights. Results are shown in Figure 28. In all plots a distinct peak appears at around 2-2.5h. This is the most common time passengers allocate to come to the airport before the flight. Very long times, well above 5h are most likely spent by flight operators (pilots, flight personnel) as those users have a well-established schedule for most days on the week by arriving at the airport very early (this can be seen in Figure 29a). It is also a bit more challenging to pinpoint the exact time passengers depart on an airplane. This depends on multiple things such as if they turned their phones to airplane mode before the flight, and if they didn't when their phone lost contact to the network and how this was processed by the baseband. When comparing the plots for domestic and international passengers, there doesn't seem to be a visible difference in retention time. Additionally, from these graphs we are not able to distinguish between users based on them travelling for business or pleasure. There doesn't seem to be a distinct number of passengers arriving earlier then 2h before their flight. We are not able to make this distinction or state that business passengers or

passengers on budget flights (which usually have a much lower number of checked in baggage) arrive at the airport much later before their flight then other passengers.

We tried to compare our results with the flight history data acquired from Zagreb airport. Two histograms were made for each data set in a certain period. An example for data collected in July of 2022 is given in Figure 29. Comparison between these histograms is difficult. The mobile network data shows the hourly aggregated time passengers arrived at the airport before their departing international flight. The second (b) graph shows the aggregated number of passengers on departing flights (one or more) for departing flights in the same hourly window. Darker blue colour indicates more passengers. For some flights a slight correlation can be noticed. An increase of passengers on the left graph is sometimes seen on the right graph with a time shift of about 2h as would be expected since most passengers arrive around 2h earlier for their flights.

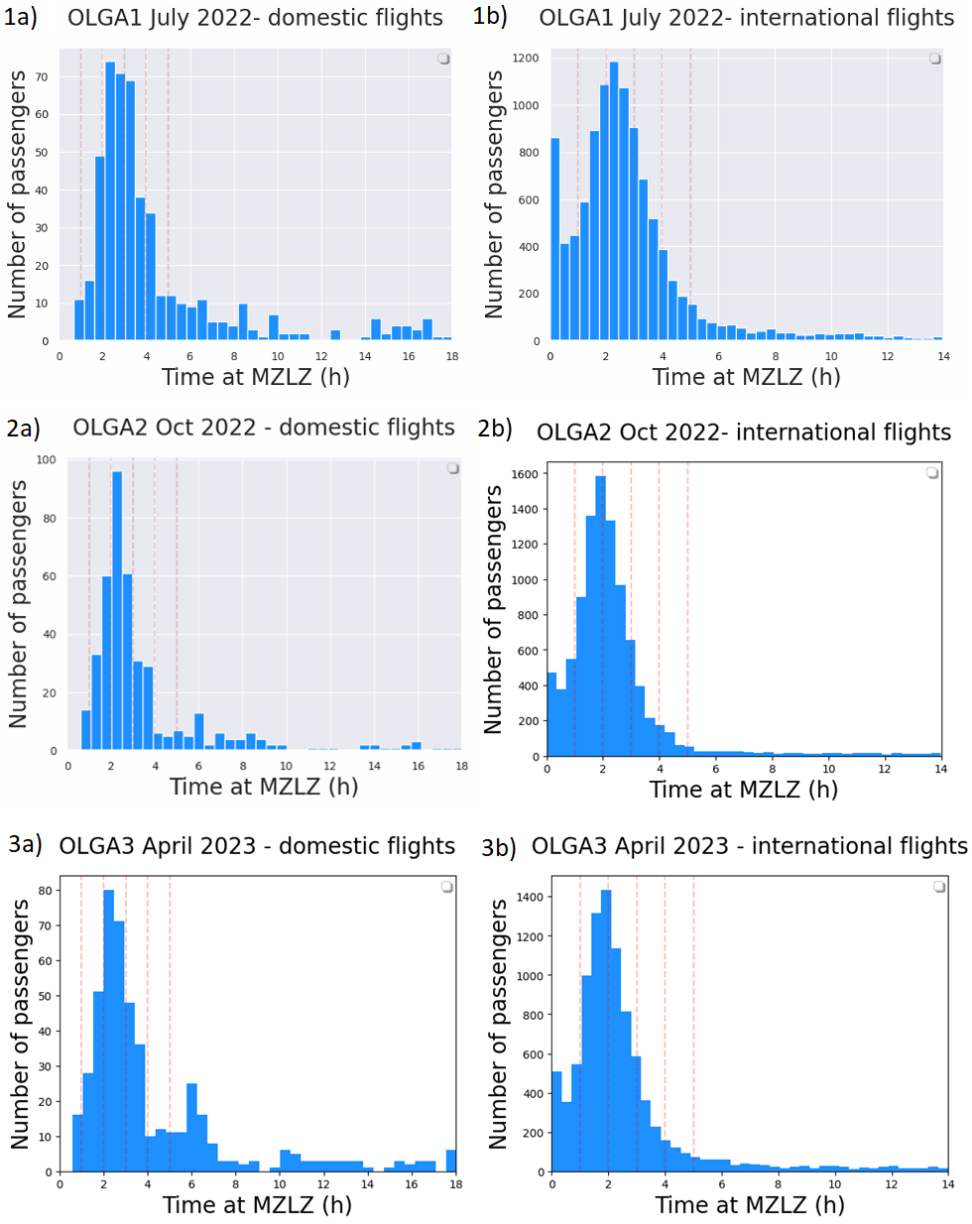

Figure 28 These plots show the total time passengers spend at Zagreb airport (MZLZ) before a departing flight. Numbers 1,2 and 3 mark a distinction between the periods mobile data was collected. Letters a and b mark a distinction between passengers leaving on a domestic or international flight.

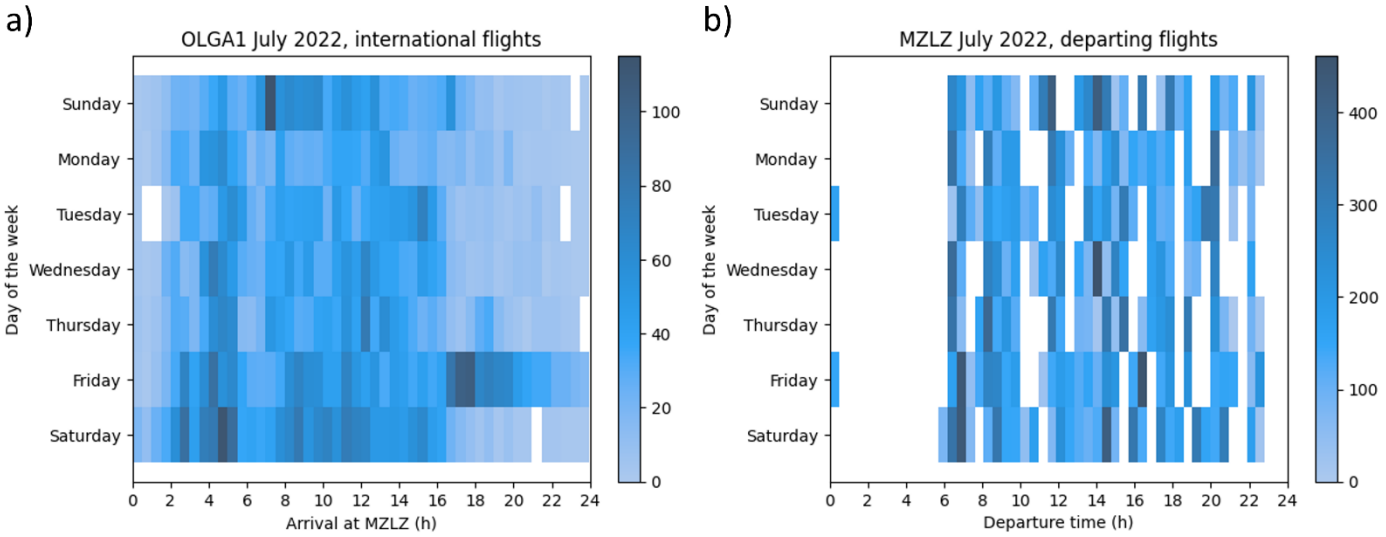

Figure 29 Comparison of the number of passengers on departing flights. a) This plot shows the number of passengers arriving at the Zagreb airport aggregated on an hourly basis during the week in July 2022. b) This plot shows the number of passengers on departing flights at the hour of departure.

The *trips* mobile data allows us to analyse routes travelled by visitors of Zagreb airport. For a use case in WP 2.1 we investigated distances crossed by users whose trips end at the Zagreb airport. The use case required insight into total distances travelled by airport users to gauge the need for more charging stations for electric vehicles at the airport. Some of our results are shown in figures: Figure 30, Figure 31, Figure 32. One can notice that most passengers travel from the Zagreb centre. A certain increase of travellers is visible for distances that can be associated with other larger cities in Croatia or the Croatian border crossing (we don't have data outside the Croatian territory, so this indicates that some airport users are likely foreigners from neighbouring countries). Data was analysed based on two snapshots of the mobile network taken during the tourist season (July 2022) and outside of it (April 2023). Additionally, we separated travellers based on their country of origin. A significant number of foreigners make their trips to the airport from the city centre.

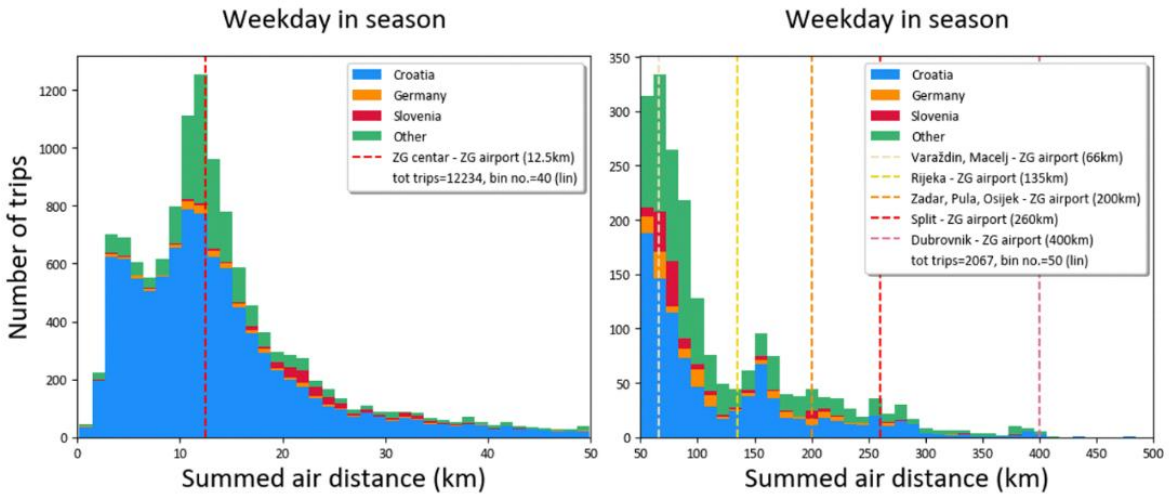

Figure 30 Summed air distance of trips with the destination sector at the Zagreb airport at a characteristic weekday in season. Left graph depicts distances bellow 50km, while the right distances above 50km.

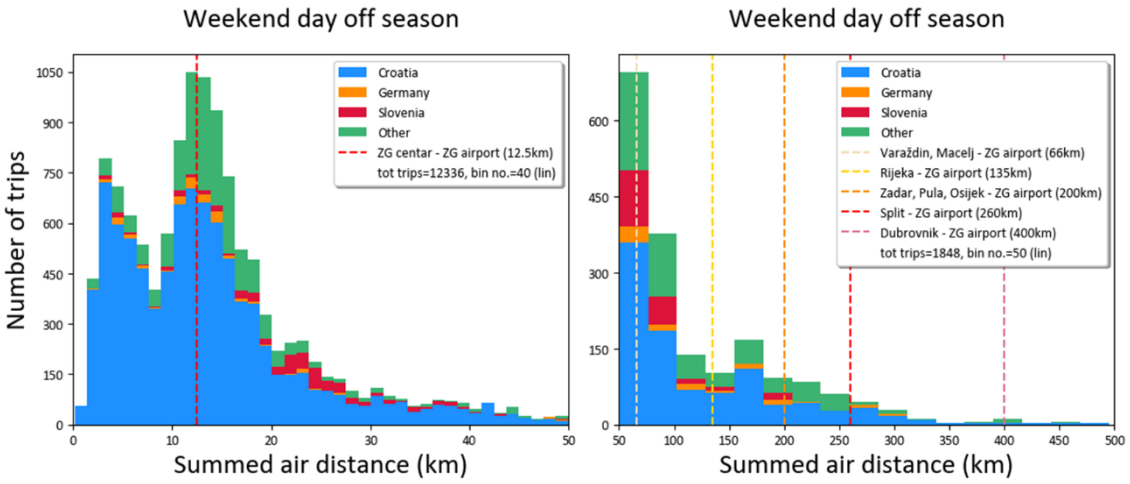

Figure 31 Summed air distance of trips with the destination sector at the Zagreb airport at a characteristic weekend off season. Left graph depicts distances bellow 50km, while the right distances above 50km.

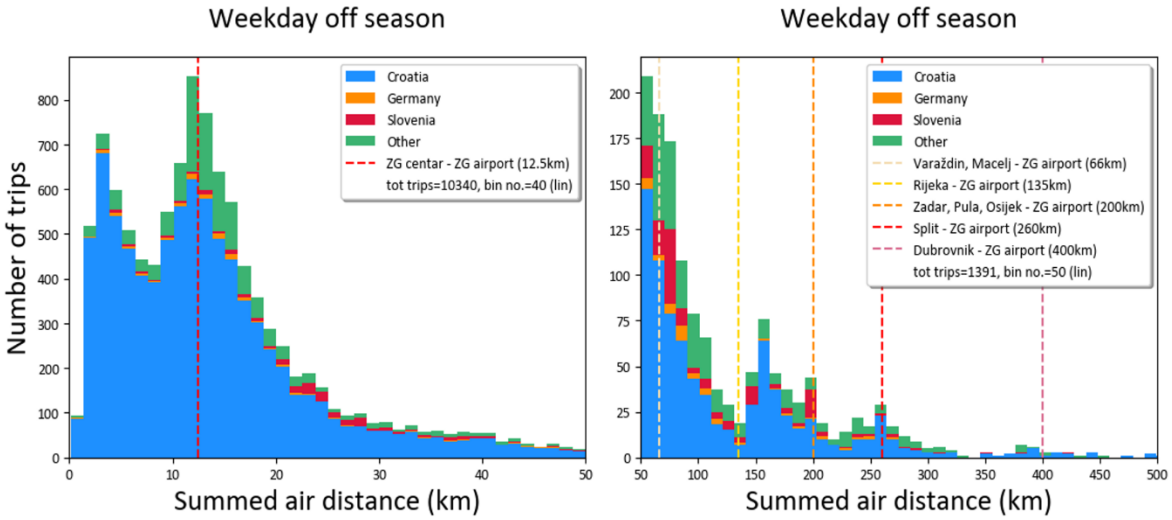

Figure 32 Summed air distance of trips with the destination sector at the Zagreb airport at a characteristic weekday off season. Left graph depicts distances bellow 50km, while the right distances above 50km.

furthermore, a comparison was made between the airport acquired passenger numbers and the passenger count estimated from telecom data. Data are shown for each day of a week in September aggregated on an hourly rate in Figure 33. For some hours of the day deviations between the two data sources differ, but overall, looking at the daily baseline, telecom data and passengers number given from Zagreb airport are similar. It should be noted that the telecom data are multiplied by 3 based on the network provider share and it acts as a good assessment of passengers' movements.

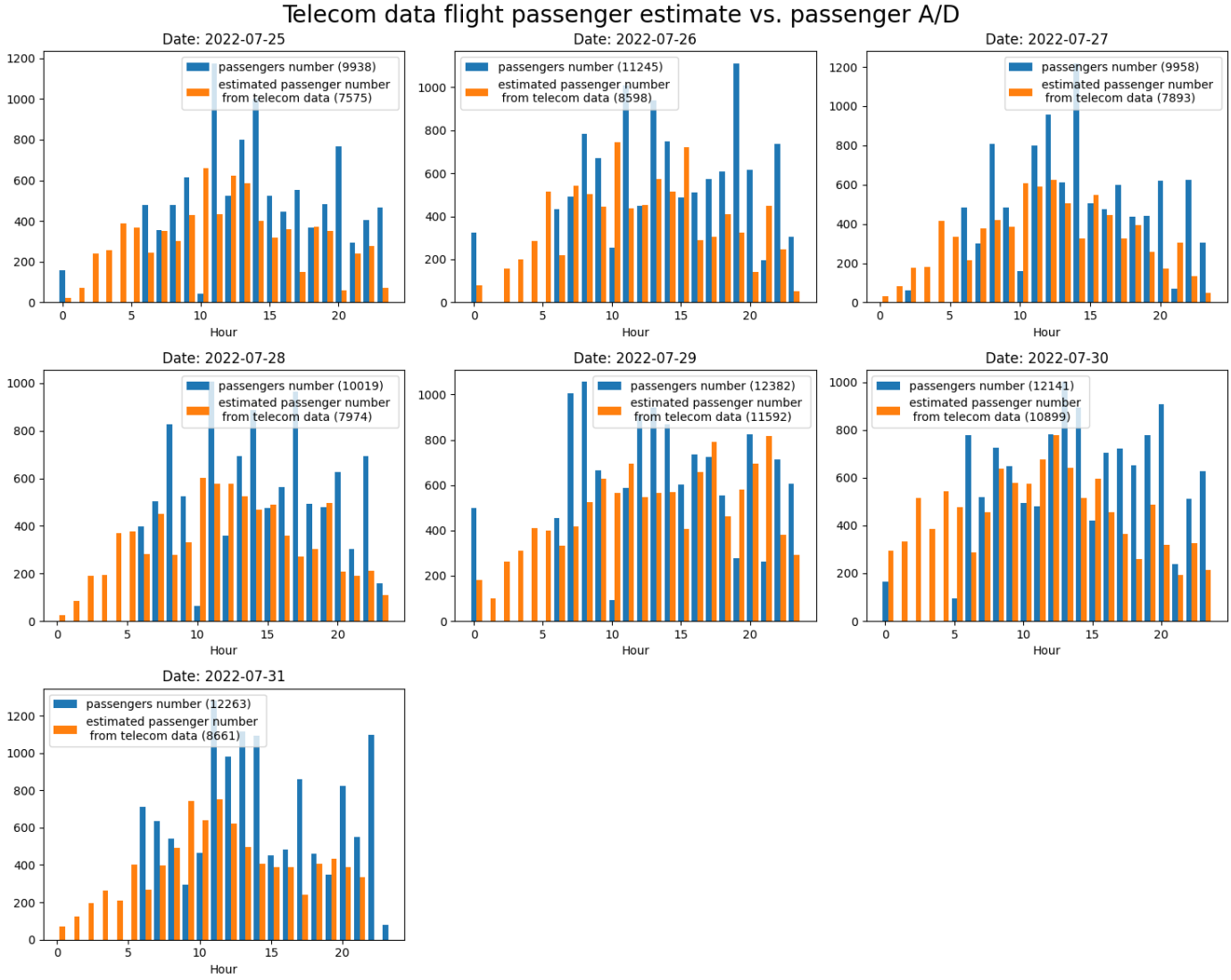

Figure 33 Comparison of the number of passengers acquired from the airport and the number of passengers calculated from the telecom data. Data is aggregated on an hourly basis and plotted for each day of the week.

4.1.5 Transport demand prediction

4.1.5.1 Prediction of usage of bus in relation to arrival of airplane (Correlation between passengers who entered the bus and passengers who landed by plane)

Information on hourly flow of passengers on the bus line 290 was acquired from a head count procedure organized by the Faculty of Transport and Traffic Sciences (University of Zagreb). This data was combined with the flight history data for comparison and a correlation check. In the diagrams shown on Figure 34 we plot results of the linear regression performed on the data. Depending on the time slice in question, either very week or a slightly negative correlation is found between the number of passengers coming to the airport on arriving flights and the number of passengers entering the bus on the airport bus station. The *p*-value is large (in range of 15-50 percent) for all cases, which indicates poor fits. Where correlation exists, it can be estimated that on average only a few percent of passengers arriving by flight use the public bus service as a means of transport towards Zagreb. This most likely indicates the reluctancy of the airport passengers to use public transport (either because of lack of information, or lack of availability), or potentially reliability issues with ZET bus line 290 passenger head count data. A more consistent head count procedure (e.g. automatized detectors) on the public bus service would alleviate all questions regarding reliability of its estimated usage.

The only available public transport system mode in vicinity of the airport is the bus service, and the very week correlation indicates towards a significant lack of its use by arriving flight passengers. This conclusion also provides a slightly deeper insight on the previously suspected prevalent use of personal means of transport and taxi service when travelling to/from Zagreb airport, which is further demonstrated in the following subchapter. Regardless of the quality of input data, we are confident the methodology described here is rigid enough for strategic planning of airline passengers public transport usage.

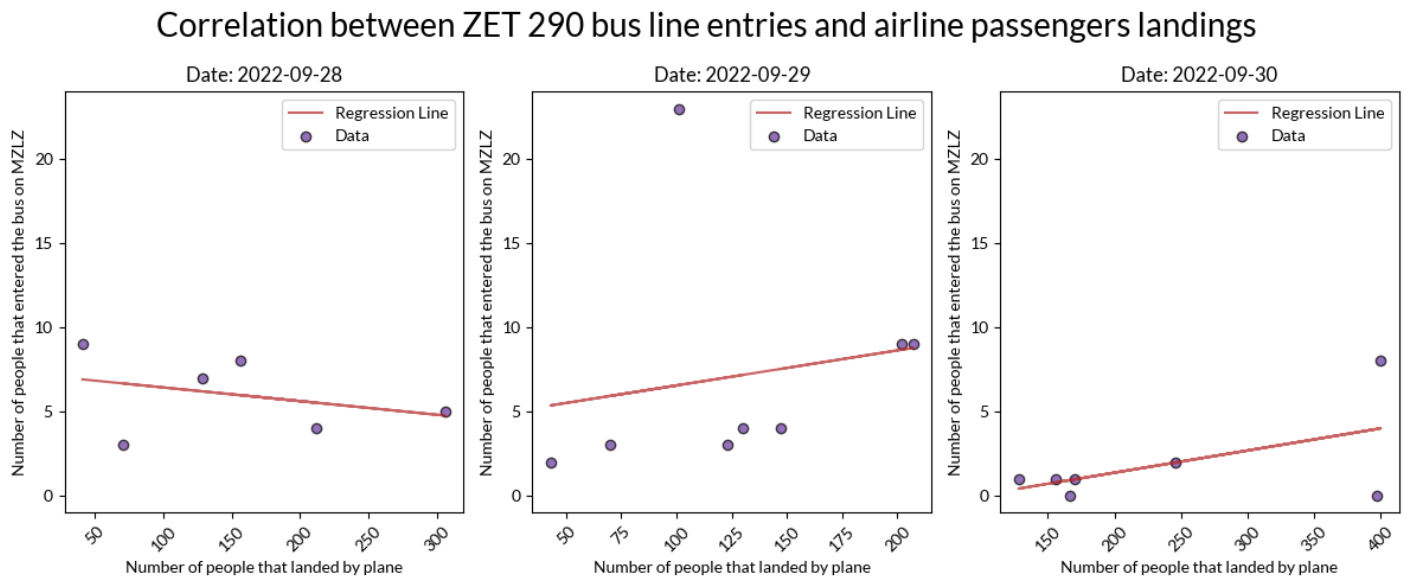

Figure 34 Correlation between passengers using public transport (by bus) and passengers arriving by plane.

4.1.5.2 Passenger transport vehicles entrance vs. passengers' A/D

Zagreb airport's parking lot vehicle flow data was paired with the number of passengers from flights history to evaluate the correlation between the two sets. This serves as a method to examine the usage of personal means of transport/taxi service by the airline users. Results are shown in the form of histograms on Figure 35 and Figure 36. The ratio between the number of passengers and vehicles was extracted to be approximately 2.2 passengers per vehicle, with an error estimate of about 15-20%. This is consistent with the previous conclusion on public transport (bus service) usage – a large portion of airport users travel to/from the airport by means of personal transport or a taxi service.

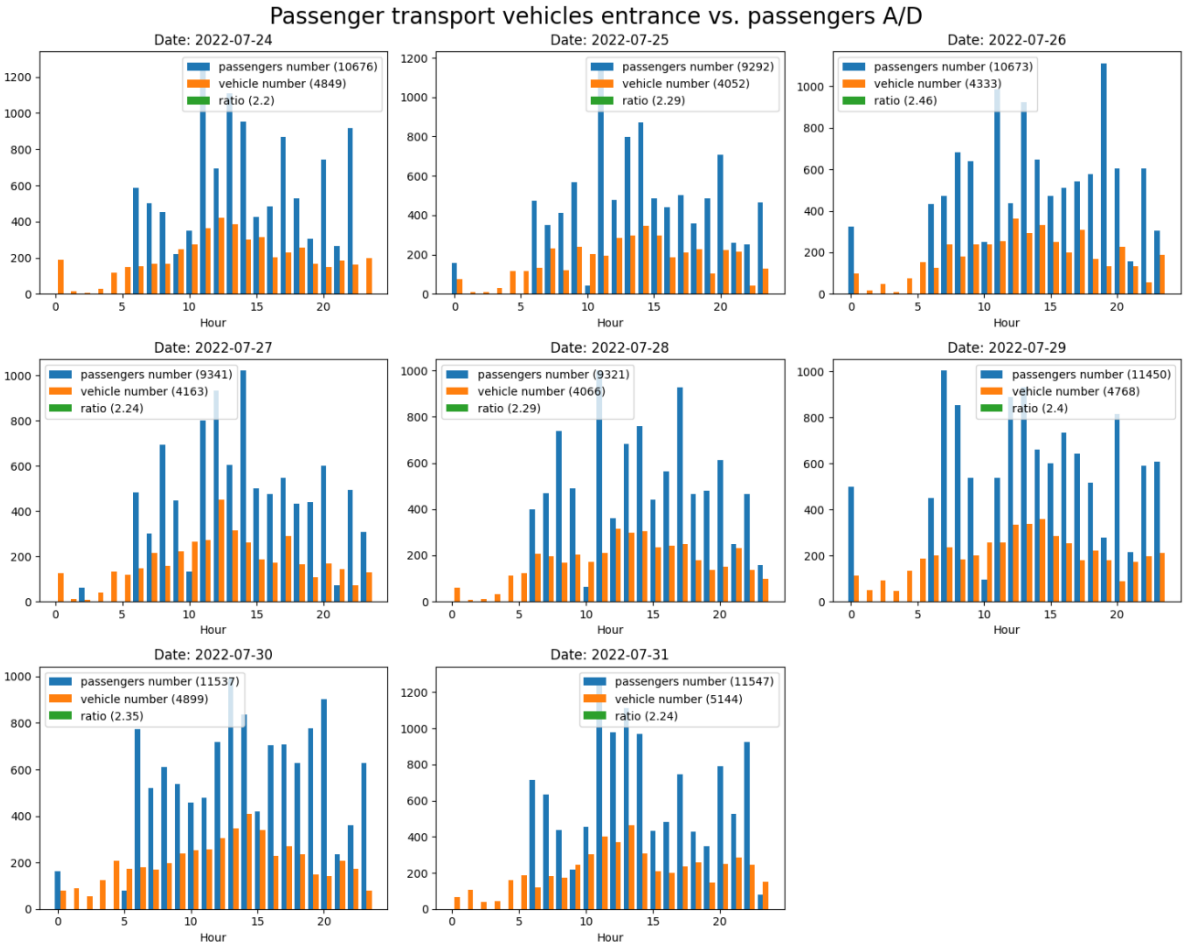

Figure 35 Passenger transport vehicles entrance vs. passengers' A/D

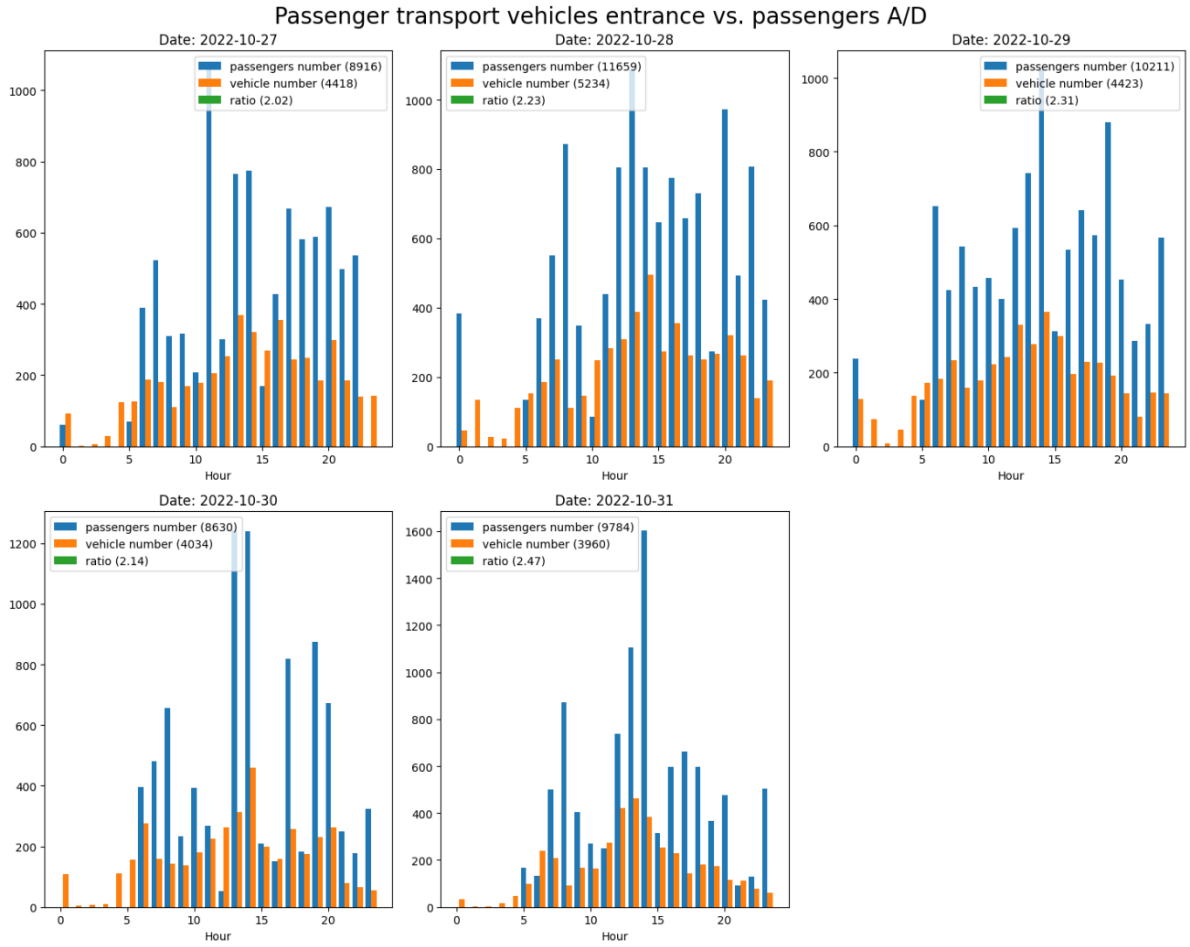

Figure 36 Passenger transport vehicles entrance vs. passengers' A/D

4.1.5.3 Employee questionnaire on transport habits

The questionnaire (described in a chapter 3.1.6) presented to airport employees included a question on preferred means of transport to and from Zagreb airport, car usage and preferred alternatives to personal means of transport to reach the airport. We also inquired into residency information of Zagreb airport employees, confirming that most of them travel to work from Velika Gorica and Zagreb.

The employees were also presented with the following multiple-choice question in the questionnaire: "If you usually drive yourself to work, which of the following commuting alternatives would you consider at least one day a week? Check all that apply." The answers given by the participants are divided into two groups. The first group consists of answers given by employees living in Zagreb and are collected and presented on a pie chart shown in Figure 37.

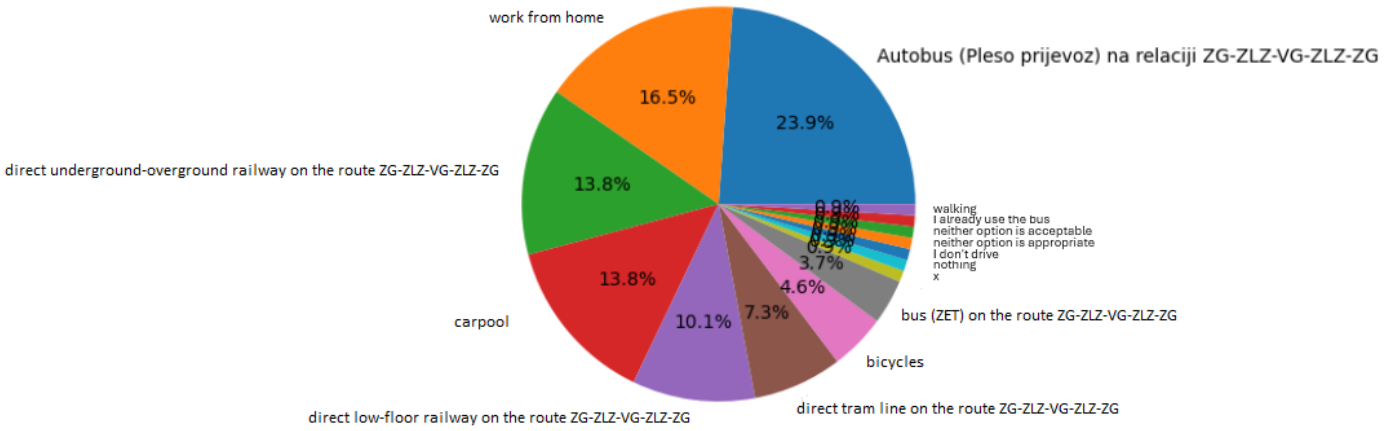

Figure 37 Preferred alternative mode of transport for employees residing in Zagreb – pie chart

The most popular employee choice (23.9%) for a preferred alternative means of travel is the Pleso bus service, which connects Zagreb airport to the Zagreb centre. We haven't been able to obtain information on the utility of the Pleso bus service which makes it difficult to assess its importance for the transportation of passengers between Zagreb and the airport. Second most popular option (16.5%) was work from home, followed by underground railway (13.8%), carpool (13.8%), tram (10.1%), bicycle (7.3%) and so on. In line with the findings of the previous two subchapters, the bus (ZET line 290) travel option was not a popular answer and only earned 3.7% of the votes.

Employees from Velika Gorica gave slightly different answers, as shown on Figure 38. As the preferred transport alternative employees from Velika Gorica chose to use the bicycle (23.5%), followed by work from home (17.3%), carpool (14.8%), going to work by foot (13.6%), public transport by the Pleso bus service (9.9%), ZET bus line 290 (7.4%), train, tram and so on.

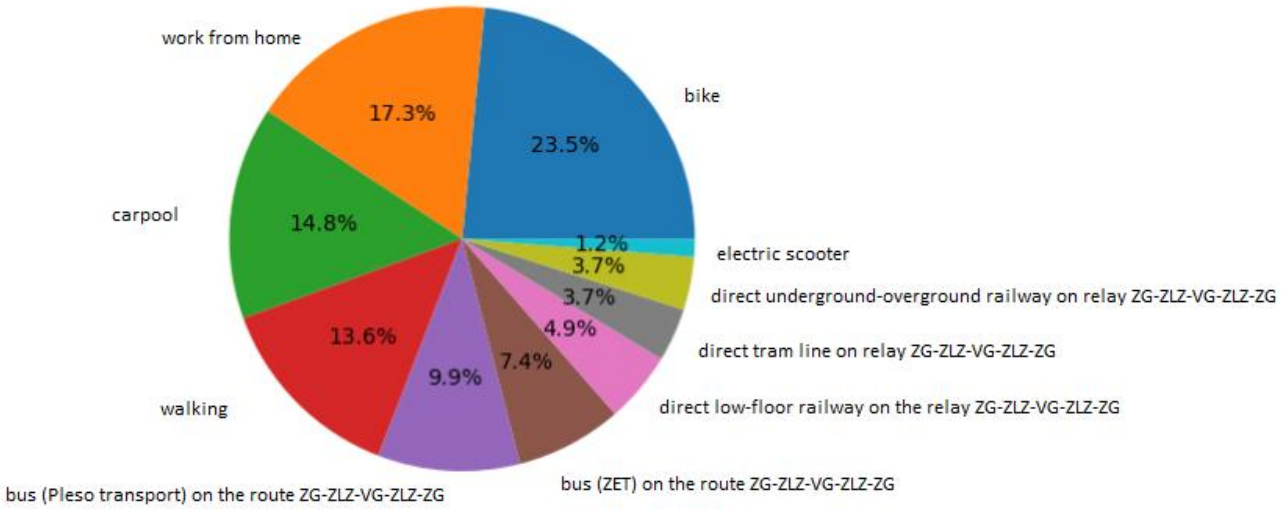

Figure 38 Preferred alternative mode of transport for employees residing in Velika Gorica - pie chart

Car users were presented with the following question: "If available, would you consider switching to an alternative, more environmentally friendly mode of transport to/from this airport?"

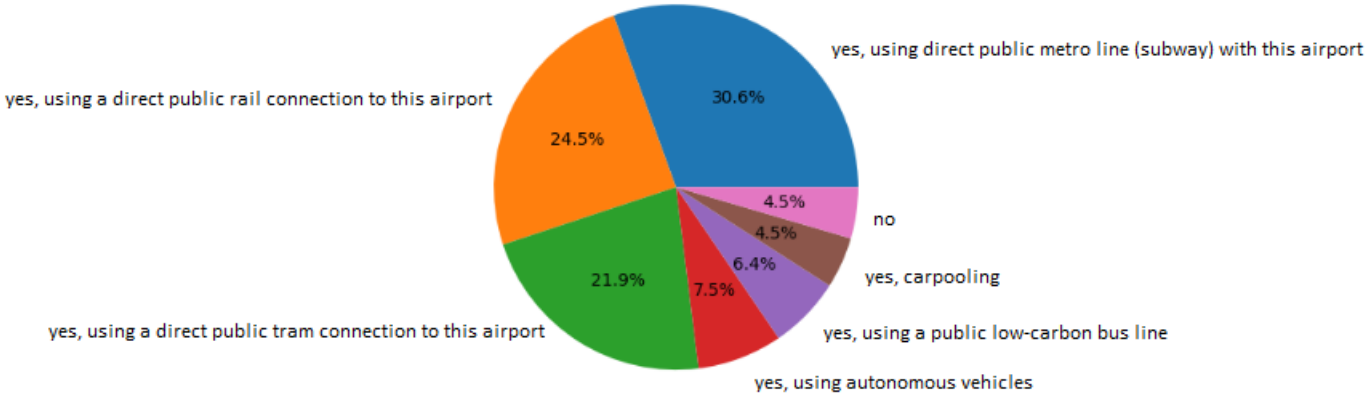

Figure 39 Preferred alternative mode of transport for employees using personal vehicles as a means of transport to/from work

Most of car users (30.6%), Figure 39, would be happy if an underground railway was available as a public transport mode, the second most popular choice was traveling by a train connection (24.5%), followed by tram (21.9%). Other choices were autonomous vehicles, bus service and carpooling. Small number of users showed no interest in changing their mode of transport.

In conclusion, employees of the airport residing in Zagreb would prefer to choose using the Pleso bus service to go to work, but not the public ZET line 290.

Nearness of Velika Gorica to Zagreb airport is reflected in the answers employees gave in the questionnaire - they are more willing to choose walking or biking to work. Employees that come to work in personal cars would, on the other hand, prefer to use the railway, train, or tram. In this case also, using the bus service is not the most favourable option (6.4%), albeit being the only currently available one.

4.1.5.4 Findings and conclusions regarding underuse of public transport

Several clues regarding the reasons for underuse of public transport were discovered while performing groundwork. One surprising finding was a quirk of the (widely used) Google maps application, which show misleading instructions for the location of the bus stop when searching for a route towards the city. The pedestrian crossing is not correctly displayed, as it is shown in Figure 40, where the red line depicts the actual shortest available and traversable pathway pedestrians usually take when walking towards the ZET bus stop, and the blue dotted line is the one suggested by Google maps. Because of this the estimated travel time towards the city centre by the Google maps application is artificially inflated to a total of 51 minutes. This value is additionally realistically inflated by the fact no direct lines are available to the city centre, and travellers must transfer from the bus service to some other mode of transport once they reach the Kvaternik Square. Unsurprisingly, the mobile network data analysis demonstrates that Kvaternik Square is rarely airline passengers' final destination, and that the Pleso bus line ends much closer to the planned destination of an average airline passenger. Additionally, existing airport signposts do not provide any directions to the ZET public transport bus stop for passengers to locate the fastest route once they exit the airport. Public transport will then more likely be avoided as a convenient means of transport by passengers travelling with heavy luggage.

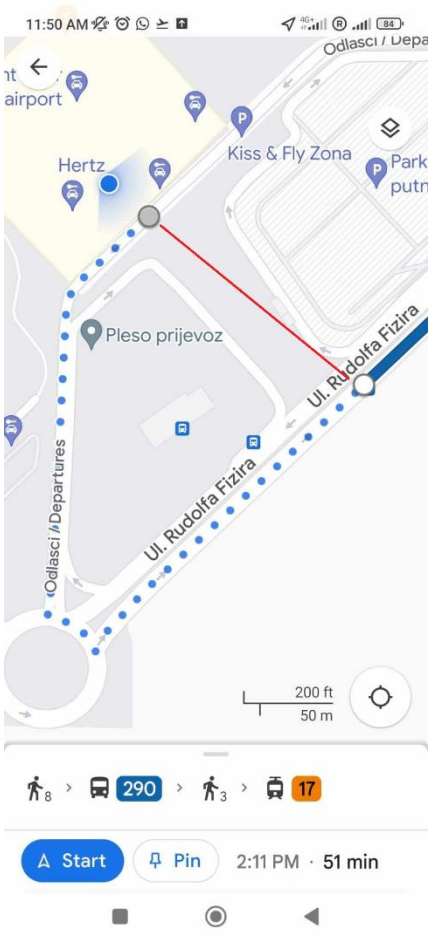

Figure 40 Google maps direction suggestion from Zagreb airport to the public bus station

4.1.5.5 Parking lot usage forecasting

Parking lots are defined by the limits of their capacity. For a traffic node with such a heavy personal vehicle/taxi service load as the Airport (as demonstrated above) this can become a strategic risk, especially considering the importance of efficiency in airport retention capacities. Any congestion that occurs in periods of high traffic load could result in airport users being at increased risk of missing flights, not to mention traffic security issues. Forecasting demand for parking space using contemporary machine learning (ML) models can serve as a strong efficiency improvement mechanism in this capacity.

The Zagreb airport parking lot entrances and exits are moderated by an automatized ramp setup maintained by a third-party company. The ramps are integrated in a system that generates information on the vehicle flow. The data is aggregated on an hourly basis and combined and correlated with flights history data (which has

also been aggregated on an hourly basis). As such it is then used for training of an ML model which forecasts the parking lot demand on a weekly interval.

On Figure 41 a graph is presented, where the blue line represents ground truth datapoints used to train the ML model, and the green line represents the parking demand forecast predicted by the ML model.

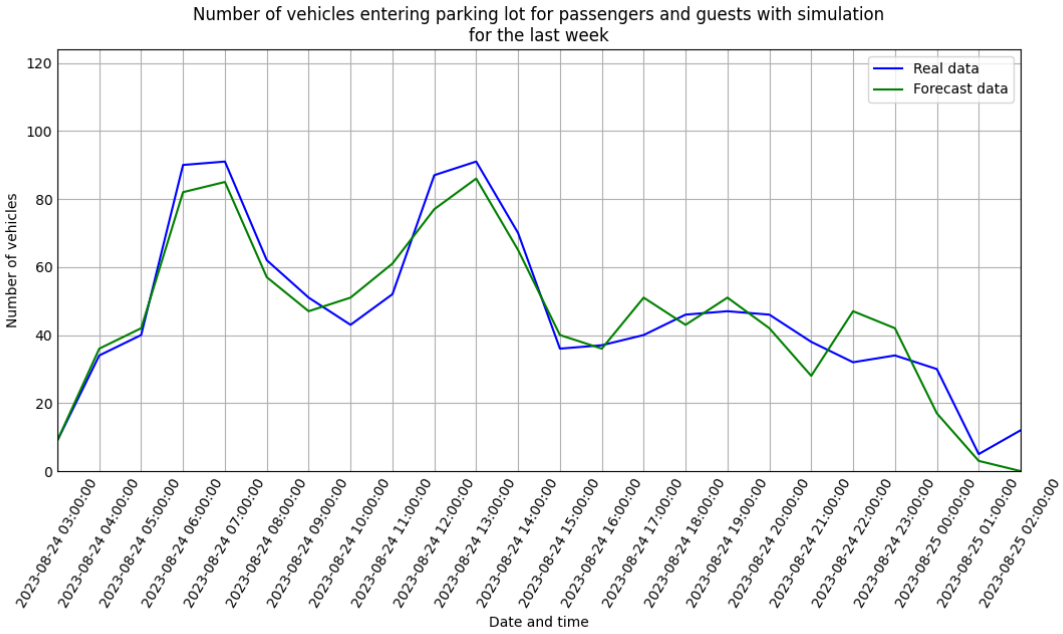

Figure 41 Comparison between the parking lot vehicle flow forecast and the ground truth data.

During the process of validation our model predicted the number of vehicles entering the parking lot with an accuracy of 80-90%, which we find to be a satisfying result. The parking lot vehicle flow data acquired from a third party is not fully reliable – many artificial double records exist, while some vehicles are not correctly registered during passage, which provides a further strain on correctly forecasting parking demands – ML models can only ever be good as the input data is. This issue was somewhat ameliorated on a more recent dataset the third party provided, which reflected on the model precision somewhat positively.

For manipulating datasets similar to the one in question, most of the tasks were performed using the Python programming language. Pre-processing, filtering, aggregation, and pre-analysis of data was performed using the popular Pandas library. Matplotlib is another popular Python plotting library employed in the data-science tasks performed. Visual insight is sometimes more intuitive than purely numerical and/or statistical analytics. For the prototype ML parking demand forecast model design LightGBM was used, yet another Python library. LightGBM is a gradient boosting framework that uses tree-based learning algorithms. In implementation phase this code has been rewritten in Java programming language in the form of an Apache Spark job.

Feature engineering is an important step in building a well-rounded ML model. Creating new features from already available data can help ML algorithms converge to a good depiction of reality, i.e. forecast more correctly.

Aside from historical parking lot usage, additional features are of temporal type, and are defined as:

- *month* – ordinal number of the month in the year.
- *day* – ordinal number of the day in the month.
- *day_of_week* – ordinal number of the day in the week.
- *hour* – the hour in the day.

Analysing the parking data shows patterns emerging on timescales with duration of a week. The cause of this pattern emergence is easily explained by the schedule of arriving and departing flights – most flights are daily or weekly. For this reason, weekly lagged features were added. Features were derived from information on traffic participants (both for arrival and departure) and the input and output traffic from the parking sensors which we aim to forecast. For example, when making a prediction for inflow traffic on the parking lot GP at the Zagreb airport, lagged features are:

- *traffic_participants_ARR_prev_week_same_hour* – feature derived from data on traffic participants on arrival flights, with a 7-day lag. This means that data from feature *traffic_participants_ARR* are shifted forward 7 days so that the data from the previous week contain timestamps of the current week.
- *traffic_participants_DEP_prev_week_same_hour* – similar to the previous lagged feature, data is shifted for traffic participants on departure flights.
- *GP_IN_prev_week_same_hour* – this feature is also derived by shifting data forward for 7 days for the total number of vehicles entering the parking lot GP within one hour.

The following graph (Figure 42) shows comparison between ground truth (red line) and forecast data (green line) for the parking lot GP at Zagreb airport.

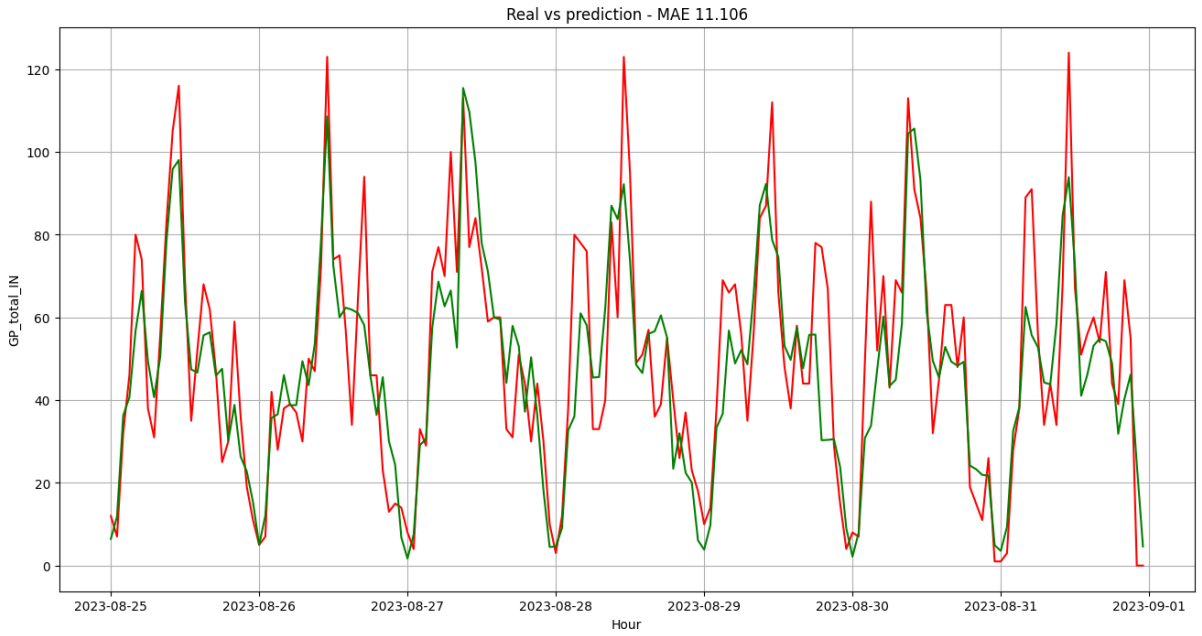

Figure 42 Comparison between real parking data and the ML predicted data.

The median absolute error is slightly above 11 vehicles per hour throughout the whole week of predicted data, which is of order of 10% of the maximum vehicle flow.

The following graph (Figure 43) lists features used for the ML model setup and displays importances of implemented features in the form of a histogram.

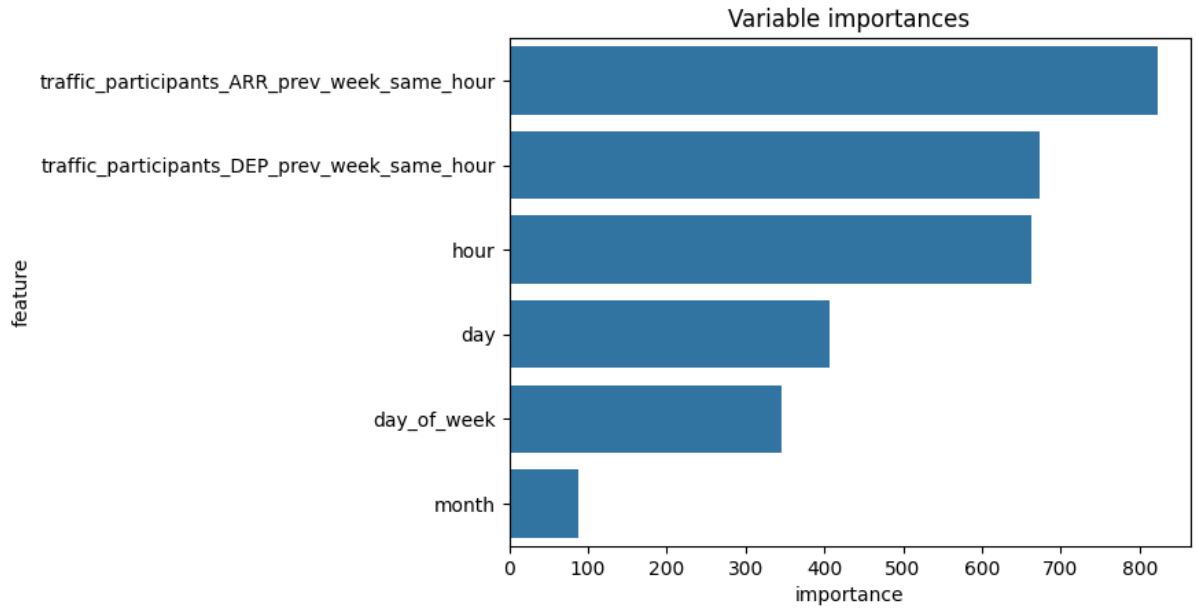

Figure 43 Features of the ML model and their importance.

Additional features were implemented in the testing phase in order to keep track of anomalies (such as holidays), but it was found either that their effect wasn't significant, or that a much larger dataset is required in order to correctly update the model (e.g. during holidays such as Christmas the parking demand plummets, but they only occur once a year, which would imply that at least a few years' worth of data is required in order to fit a model that predicts parking demand during Christmas correctly). The fact that with only a few features a model can predict parking demand very well is an assurance into prospects for an efficient and useful implementation of such a technology.

In summary, prototyping of the ML model was performed using Python within its libraries. Before deployment everything had to be converted into an Apache Spark job. The data management platform design required for the Apache Spark jobs to be written in Java. This allowed for some parts of the code needed to read data from the data management platform and to store data into the platform to be reused.

4.1.5.6 Research into machine-learning based virtual traffic counters, and a test implementation

Research was performed into prospects of an innovative real-time traffic monitoring application employing Performance Management (PM) data available from the mobile TC network, based on [9]. A preliminary test case study was performed by choosing one of the more trafficked state roads connecting the city of Zagreb to the airport (in this case - state road D30). Aside for this connection condition, the road was chosen according to two additional requirements: a ground truth data source is necessary on which to train ML models appropriately, and a source of PM data stemming from network infrastructure adjacent to, and oriented towards the road is necessary in order to represent (at least partially) the weight of mobile network activity of the users traversing the road. Ground truth traffic flow data (collected by induction-based physical traffic counters) is made available through the Croatian National Access Point website (https://www.promet-info.hr/en) and is collected to an internal PostgreSQL database by pinging the website APIs through a set of automatized Python-based scripts. The traffic counters proved to be a powerful and reliable source of ground truth, and their presence provides a very useful insight into traffic densities surrounding the airport on its own., The network PM data is available to Ericsson employees internally.

The topology of the setup is available on Figure 44, with the physical traffic counter location marked by the blue dashed circle, and the idealized mobile TC antenna coverage is depicted by a pink circle sector. The state road D30 is marked with a bold purple dashed line. It is easily noticed that the topology setup is far from ideal – the models are expected to predict real traffic flow through a traffic point separated from the part of road covered by the mobile TC antenna by ~3 km, a distance which contains several traffic drains when traversed by car along the D30 state road.

A set of features is chosen, composing of a single feature representing the mobile TC network usage weight, and a few temporal features. The feature describing the mobile TC network is aptly named *users* and is a linear

combination of several PM variables. Several models are employed, e.g. Random Forest, Histogram Gradient Boost, etc. An example of a comparison between the model predictions and reality is shown in Figure 45 and Figure 46. It is obvious that these kinds of *virtual traffic counters* show great potential as a technology of the future – even for this evidently suboptimal network topology setup and minimal effort in model tweaking, the accuracy score for the model can be improved by 20-40% by using the mobile TC network PM datasets, which (as authors in also correctly note, [9]) are not plagued by privacy and scaling issues. Ingesting a more encompassing dataset (e.g. all the available PM data surrounding the airport), performing the training on a larger set of ground truth data, and investing larger effort into feature engineering and model building is expected to greatly improve model accuracy.

Figure 44 Mobile network/road network topology interplay for the case study of D30

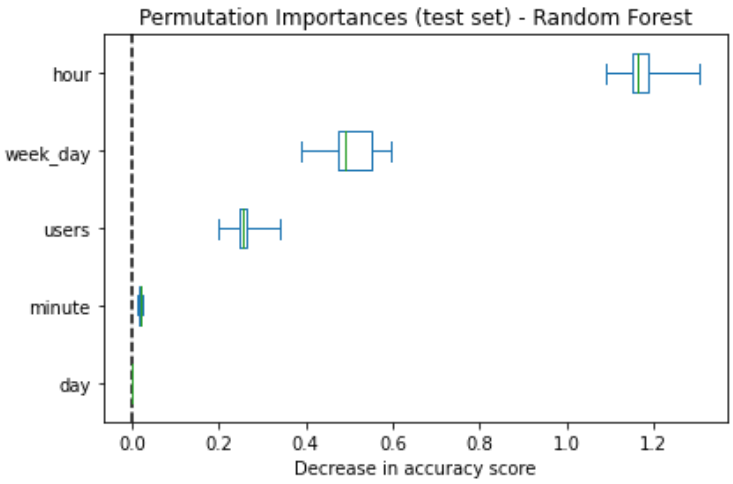

Figure 45 Permutation importances on the test dataset for the Random Forest model

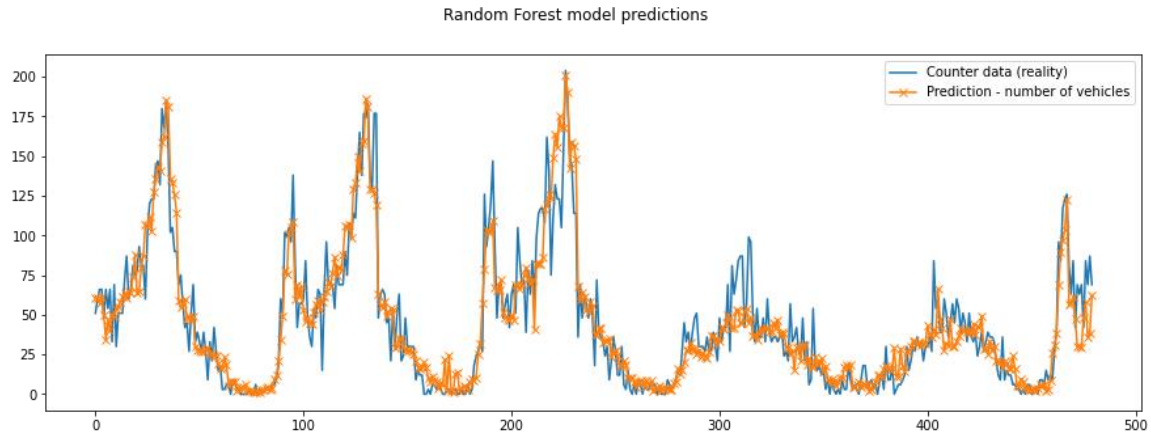

Figure 46 Comparison between model predictions and reality for the test dataset; x-axis represents unfolded temporal datapoints

4.2 Paris Airport Charles de Gaulle

Newly developed IT solution (platform) was used for traffic flow optimisation using identified use cases developed for an extensive analysis in Zagreb Airport area. For the case of Zagreb Airport all required data types (Transport data, public transport data, airport data, parking data & other data) were available and therefore, all identified use cases were supported. However, the tool was designed in a way that if a certain area has only a limited data set available, it can support use cases that rely only on those available data sets. And therefore, to demonstrate the applicability of the tool on a limited data set, an analysis has been conducted for the Paris Airport Charles de Gaulle. Paris Airport Charles de Gaulle for this analysis made available flight data with the number of passengers per flight and parking data. Data was collected for the period between February and March 2024. Based on the available data sets, analytical use case regarding transport demand analysis was performed.

Desktop analysis has shown that there are several public transport options connecting CDG. Some of them are a Shuttle to Parc Asterix, a bus RoissyBus that goes to the Opera in Paris, a train RER (Paris by train) that passes near airport Orly, Bus line 350 that ends in 'Gare de l'Est' Paris, bus 351 with the route ending at the 'Nation' in Paris, and a Shuttle to Disneyland. The public transport map is shown in Figure 47. Passengers have 6 different connections to the airport. We explored the frequency of these available public transport options.

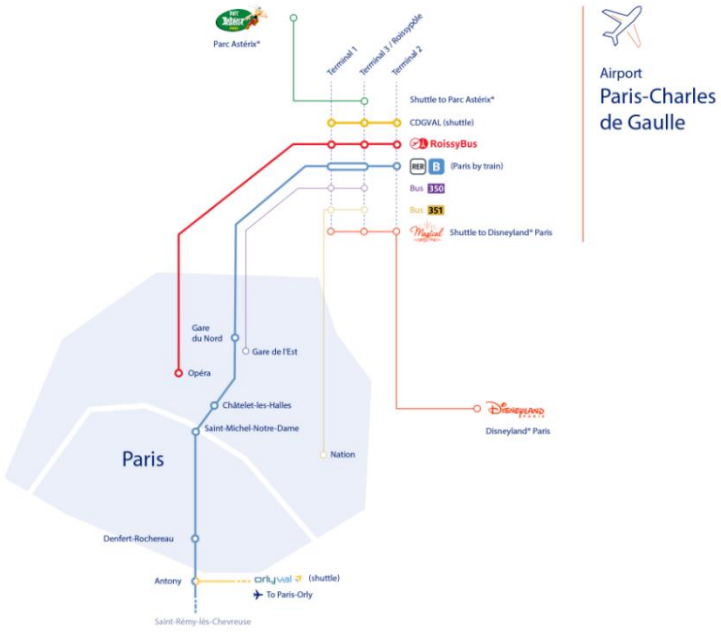

Figure 47 Public transport Charles de Gaulle

Most frequent available buses are by the operator Roissy Bus, leaving every 15min during the day, and every 20min during the night. The RATP line 350 departs also every 15-20min during the day, and every 35min in the evening. The RATP line 351 offers a frequency of 30min during the day and night. Paris RER connects CDG to the Paris centre with trains departing every 15min.

This data reveals that the average frequency of offered public transport from CDG is 14 lines per hour, i.e. a passenger can depart to or from CDG every 4.28 minutes. Frequency at night is lower than during the day. There are 9 public transport lines per hour available to passengers. This means on average a passenger can take public transport to or of from CDG every 6.67 minutes.

Since the alternative to the public transport services might be utilisation of personal vehicles, following analysis will be focused on parking data in combination with passenger arrival and departure data.

The following figure (Figure 48) shows the ratio of vehicle entrance and passengers' arrival and departure. The average calculated ratio is around 8. This indicates that the occupancy of vehicles is around 8 persons when arriving or departing to/from the CDG airport, or to be more precise, 8 passengers arrive on CGD airport by public transport compared to one arriving by personal vehicle. Such a large ratio value indicates the importance that public transport has in traveling to this airport.

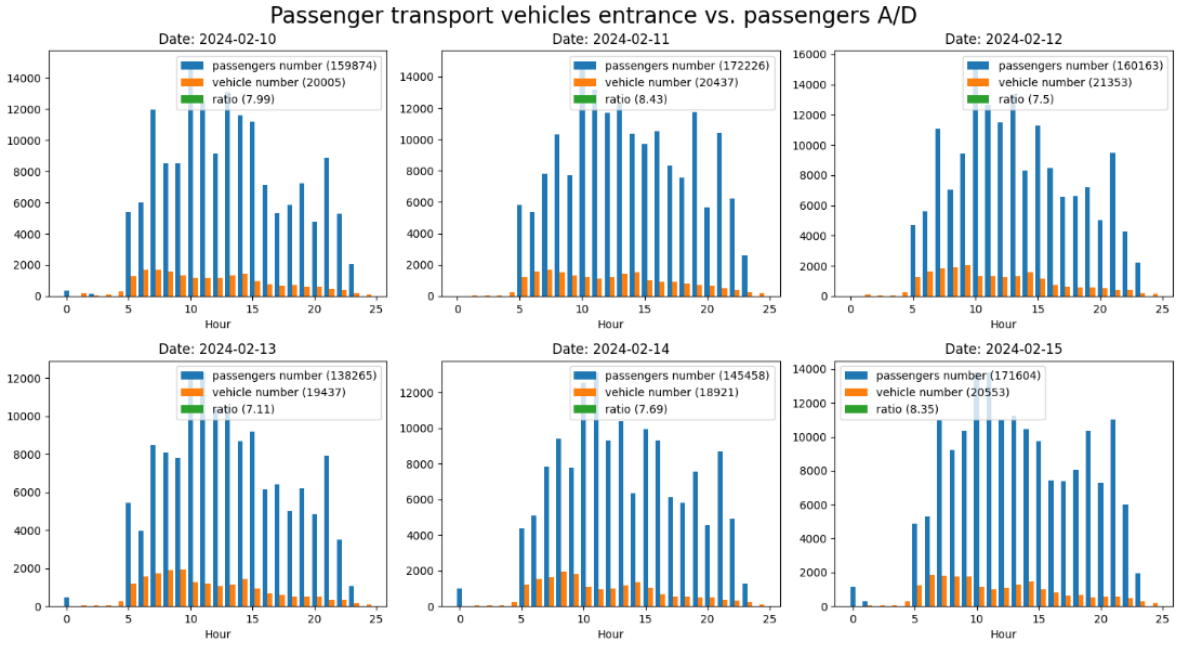

Figure 48 Passenger transport vehicles entrance vs. passengers A/D on CDG

It can be concluded that the public transport offer is appropriate and corresponds to transport demand at CDG. Users of public transport are offered with a frequent and convenient way to travel to the airport, and the public transport is the dominating transport mode. Ratio of 8:1 for passengers arriving by public transport vs. arriving by personal vehicle proves this thesis and proves that appropriate public transport offer is the reason for a lower tendency for passengers to rent a car of take a taxi to reach this airport.

For other CDG use-cases telecom data is needed but it is not available. Also, a more visual representation of the analysis in not possible because of a lack of data (number of passengers entering/exiting public buses). Various data sources would be needed for a more detailed analysis.

Future research will be focused on inclusion of new data sets in form of data alternatives for analytical use cases where initial planned data sources were not available.

5 Key Performance Indicators (KPI)

In the context of the H2020 project OLGA, a proposal had been put forward to develop an IT solution aimed at optimizing multimodal traffic by integrating data from three distinct sources. The proposed analytical use case seeks to identify and enhance public transport connections to the Zagreb airport (Velika Gorica) with the objective of benefiting local residents, airport passengers, and employees. [10] As part of the project's evaluation framework, key performance indicators have been outlined to gauge its effectiveness, including an assessment of the environmental impact associated with the expansion of the multimodal transport network. KPI for project OLGA was defined in the document: OLGA KPI definition: master document WP1.1.

5.1 The extension of the network and its maximum capacity

First two key performance indicators that are established: the extension of the network (1) and its maximum capacity (2). The linear length of the network serves as a reliable metric for its size as the area it insists on can be limited by geographical factors or the size of the considered airport. Linear length can be easily obtained through direct means or simply using tools available online. The maximum capacity is contingent upon the number of active vehicles within the network and the value can be retrieved from information given by network operators or through direct observation. [12]

$$\text{MS1.1: } \mathrm{Ext} = \mathrm{Length}\ [\mathrm{km}] \tag{1}$$

$$\text{MS1.2: } \mathrm{Cap_{max}} = \mathrm{N_{vehiclesmax}} * \mathrm{Cap_{vehicle}}\ [\mathrm{pax/h}] \tag{2}$$

An analysis is carried out on points of interconnection between processes and the possible realization of identified use cases, leading to evaluation of determined KPI's and generation of required outputs.

Figure 49 outlines the methodology proposed for the realization of public transport optimization based on dedicated stakeholder input and the calculations for the corresponding KPI's. The process begins with an analysis of the current public transport offerings, drawing from data sources such as public transport stops, lines, and timetables provided by public transport operators in standardized formats like GTFS or NeTEx. Additionally, data on vehicle characteristics, including type and capacity, is required for KPI calculation.

By following the methodology, we can delineate the spatial areas (sectors, locations) served by current public transport lines that cater to the airport and compute the length and capacity of the network. Furthermore, leveraging anonymized telecom data sets containing the origin-destination matrix and user type data allows for an analysis of transport demand. The origin-destination matrix provides spatial information, detailing transitions between sectors and establishing pertinent routes. This facilitates the identification of common

sectors where users originate, with destinations in the airport sector, along with the transport modes used between them.

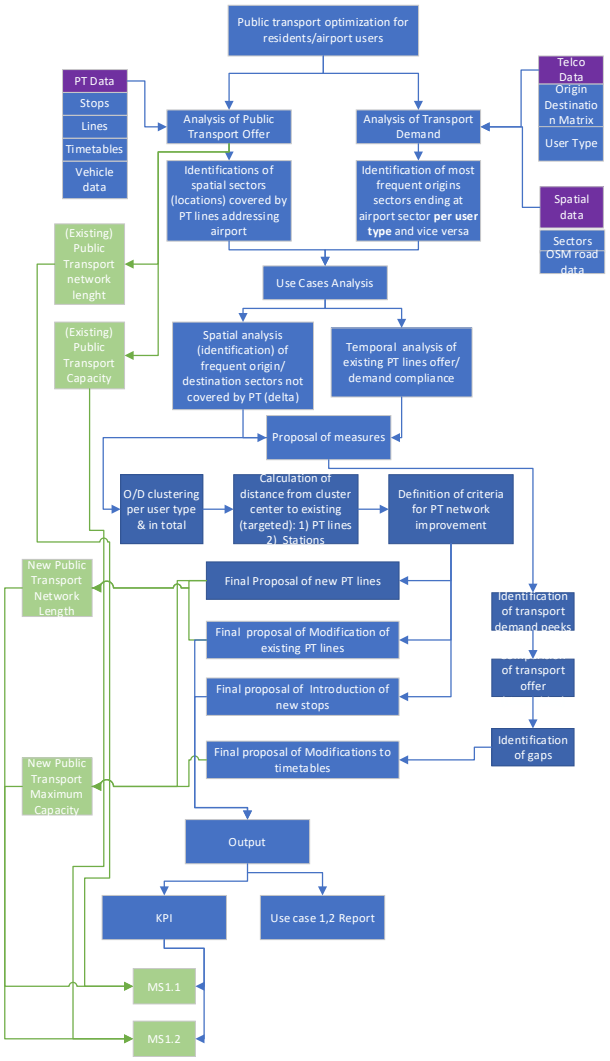

Figure 49 Proposed methodology for analytical use case execution in relation with calculation of KPIs MS1.1 & MS2.2.

Once the public transport supply has been determined and analysed, we can address the use case.

By identifying frequently used origin/destination sectors not covered by public transport (delta) and conducting a temporal analysis of existing public transport lines (supply/demand), compliance appropriate measures can be recommended. The objective is to establish origin-destination clustering per user type and in

total. Additionally, a detailed assessment of distances between origin and destination cluster centres is carried out, taking into account the existing public transport lines and stations. Criteria for public transport planning, such as the expected walking distance from public transport stations and interconnection with other transport modes and infrastructure, is used to define a preliminary set of measures and propose new public transport lines to connect currently unlinked areas. Proposed modifications to the current locations of public transport stations and lines aim to enhance existing connectivity and infrastructure. A proposal for adjusting timetables based on peak transport demand and a comparison with the current offerings can now be provided. Final measures will be structured following validation by key stakeholders, including city authorities and public transport operators. Upon the implementation of the proposed measures, the values required for KPI calculation will be determined, encompassing new parameters for public transport length and maximum capacity.

5.2 Traffic around airport

Next KPI (SOC3) is named “Traffic around airport”. This KPI gives a measure of vehicles circulating (entering/exiting) through the airport on an hourly basis. Calculation of this KPI is based on equation (3). The project's methodology is based on utilizing traffic data recorded at the outskirts of the airport area or data collected from a series of cameras positioned along the main roads leading to the airport. This assessment considers both public and private vehicles.

$$\text{SOC3:} \qquad \mathrm{NV}[-/\mathrm{hour}]; \qquad \Delta\mathrm{nv} = \frac{\mathrm{NV_{meas}} - \mathrm{NV_{ref}}}{\mathrm{NV_{ref}}} \qquad (3)$$

In the envisioned IT solution concept, the following methodology is outlined, as depicted in Figure 50.

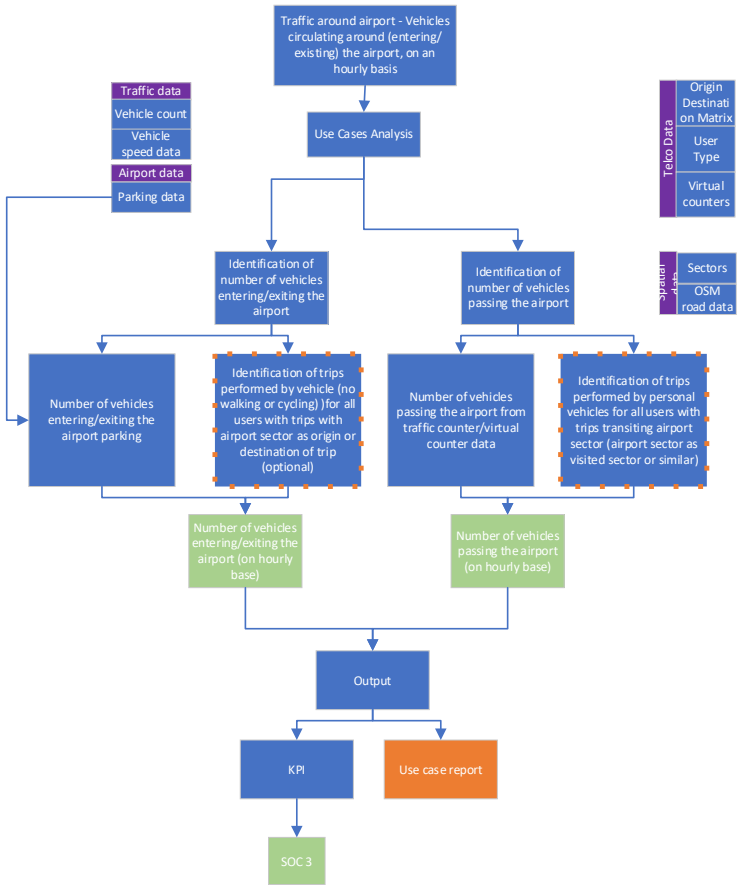

Figure 50 Proposed methodology for analytical use case execution in relation with calculation of KPI SOC3.

To assess the KPI for traffic around the airport, specifically the number of vehicles circulating (entering/exiting) the airport on an hourly basis, we utilize several data sources: traffic data (vehicle count and speed), airport data (parking), and anonymized telecom data (origin-destination matrix, user types, virtual traffic counters). These identified data sources are employed for two purposes: a) identifying the count of vehicles entering/exiting the airport, and b) identifying the count of vehicles passing the airport.

For a), we identify vehicle trips (excluding walking or cycling) for all users whose journeys involve the airport sector as the origin or destination (optional). This data is then correlated with the count of vehicles entering/exiting the airport parking. For b), we identify trips made by personal vehicles for all users who have visited the airport sector in transit (airport sector as the visited sector or similar). This data is then paired with the identified count of vehicles passing the airport as detected by traffic or virtual counters.

Ultimately, to calculate the SOC3 KPI, we use the identified counts of vehicles entering/exiting the airport and the number of vehicles passing the airport sector on an hourly basis.

5.3 Greenhouse gasses (GHGs) from fuels combustion

Next KPI is called “GHGs from fuels combustion (means of transport, thermal power plants)”. Using information on the fuel consumption of thermal power plants and other various means of transport that use combustion engines, one can calculate the emissions factors for any period for which consumption data are available. By comparing between emissions indicators, before and after implementing specific actions, we can quantify the obtained results.

CO2 (or CO2e) emissions from fuels, can also be quantified as stated in EU ETS:

$$[M]_{fuel\ consumption}[kg];\ \text{Net Calorific Value}_{fuel}[\text{kWh}/kg];\ EF_{components}[kg\text{CO2e}/\text{kWh}] \quad (4)$$

$$M_{GHG\ Elec}\ [kg\text{CO2e}] = \sum\ (M_{fuel\ consumption} \times \text{Net Calorific Value}_{fuel} \times EF_{components}) \quad \mathbf{(5)}$$

The standard parameters are published yearly by the national authorities of the EU member states and can be used for the calculation. Usually, the ETS emission factors consider only the CO2 released during the fuel use phase. In ISO 14064-1:2018 these are “Category 1” emissions (Direct GHG emissions). In a life cycle approach, the emission contributions related to upstream emissions arising from fuel generation and fuel transportation/distribution could be also considered; in ISO 14064-1:2018 these are “Category 3” emissions (Indirect GHG emissions); possible source for “upstream” emission factors for fuels: ECOINVENT. Biofuels should be considered as such only if they meet the requirements of the relevant European Regulations. In the envisioned IT solution concept, the two approaches are possible, as depicted in figures Figure 51 and Figure 52.

Calculation of the KPI GHG2 (“GHGs from fuel combustion (means of transport, thermal power plants)”) is approached by first analysing the use cases "Identification of all airport related migrations (airport sector is origin or destination of trips)". Primary data source for this analysis is Telco data (Origin - Destination Matrix, User Type, Virtual Counters), followed by Spatial data (Sectors and OSM road data), Public Transport Data (Timetables and Vehicle data), Traffic data (vehicle count), Airport data (parking) and statistical data (vehicle per engine type, average consumption per personal vehicle, average consumption per public transport vehicle). The use case analysis begins by first identifying the most probable transport mode for all related migrations, then followed by a calculation of the total distance travelled per transport mode.

An estimation of fuel consumption for trips performed by personal vehicles was given by calculating the average speed of vehicle users from telecom data (optional) and by taking an estimate of fuel consumption for personal vehicles per engine type (petrol, diesel, electric…). The latter was based on statistical data for all trips

in the targeted time frame. An estimation of fuel consumption for trips performed by public transport will be calculated using information on fuel consumption for public transport vehicles based on timetable data and vehicle (type) data for all public transport departures during a targeted time period and, if available using an estimate of average speed from statistical or real time data. Based on these estimates of GHG values and by using specialised open-source SW (i.e. QGIS with plugins), the output for terms of KPI GHG2 can be acquired.

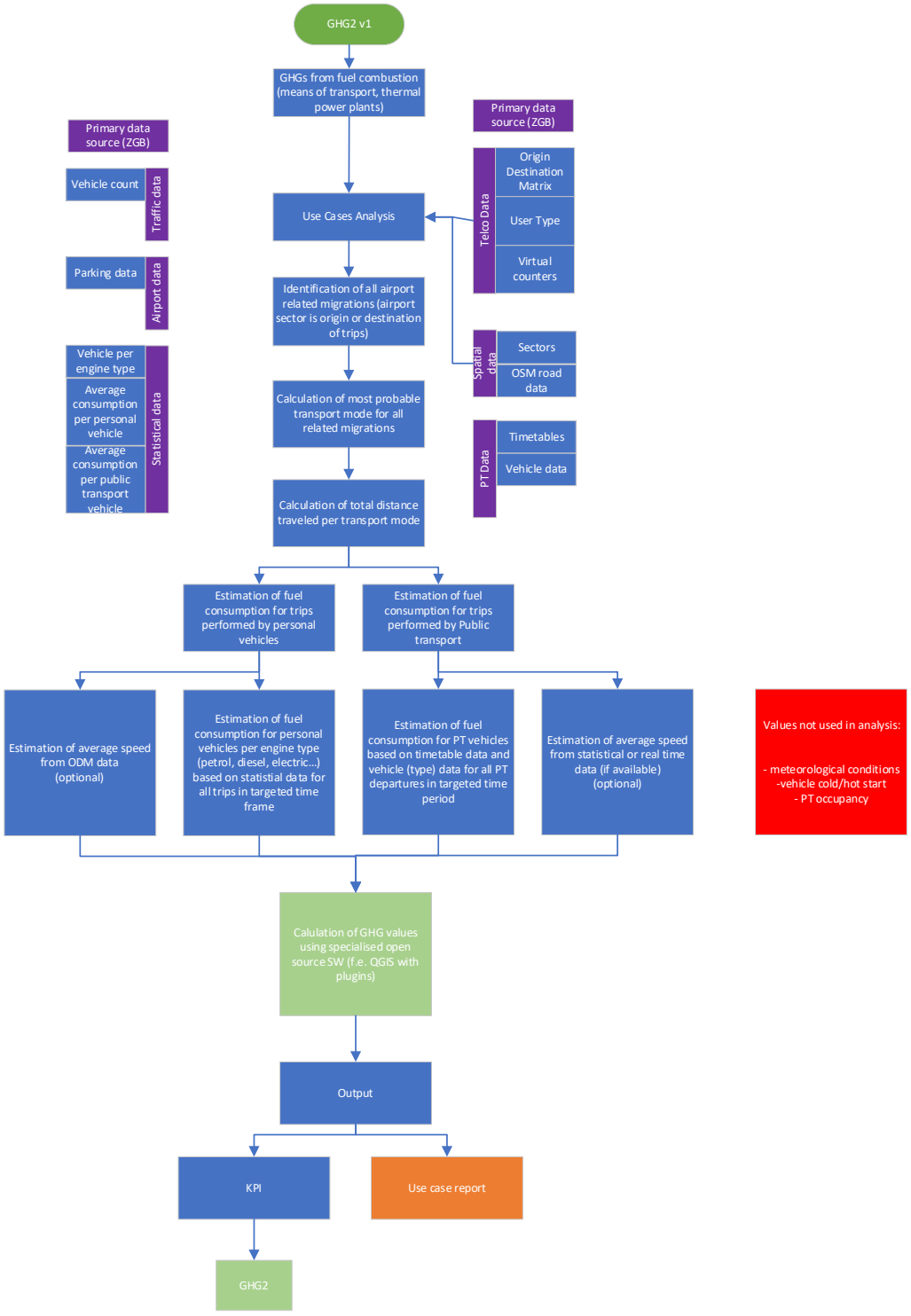

Figure 51 Proposed methodology for analytical use case execution in relation with calculation of KPI GHG – approach 1

Second approach for calculating KPI GHG2 (GHGs from fuel combustion (means of transport, thermal power plants) utilises following data. Primary data source for this analysis is Telco data (Origin - Destination Matrix, User Type, Virtual Counters), then Spatial data (Sectors and OSM road data), Public Transport Data (Timetables and Vehicle data), Traffic data (vehicle count), Airport data (parking) and statistical data (vehicle per engine type, average consumption per personal vehicle, average consumption per public transport vehicle). This KPI requires identifying all airport related migrations (airport sector is origin or destination of trips), calculating the most probable transport mode for all related migrations. The latter can be addressed when one calculates the total distance travelled per transport mode. The vehicle count data per vehicle type from traffic counter data (or virtual counter data) is needed along with a categorisation of vehicles per type (personal vehicles, busses, trucks…). Results of these analyses are takes as input data to give an estimation on vehicle fuel/engine type and average consumption based on statistical data. This input, together with the identified airport perimeters (length of roads within the airport perimeter) will be used to calculate the total distance travelled per transport mode and fuel/engine type. This data, combined with estimation of fuel consumption per transport mode and fuel/engine type will be used for calculation of GHG values using specialised open-source SW (e.g. QGIS with plugins) and requires KPI's.

Figure 52 Proposed methodology for analytical use case execution in relation with calculation of KPI GHG – approach 2

6 Conclusion and outlook

During this project, an IT solution was developed as a decision support platform for optimizing traffic flow in airport gravitational areas. By following the FRAME architecture, relevant stakeholders were identified and involved in the process of drafting requirements for analytical tools. The primary objective of the project was to establish a platform that would be genuinely utilized by stakeholders, providing them with added value in the context of data-driven operational and strategic decision-making. The analytical solution was constructed using available data sets, encompassing transport data, public transport data, anonymized telecom mobile network data, airport data, parking data, and other sources. A data aggregation platform was utilized to aggregate and manage all available data sets required for analytical use cases, employing EU-proposed and defined data formats, standards, and protocols where possible to ensure the interoperability and reusability of the solution. Diverse use cases were developed to demonstrate the capabilities of the tools to various types of stakeholders, including municipalities, public transport operators, airports, and airline companies, etc.

Several use cases were developed within the project:

1. "Public transport optimization for nearby residents and airport users (both passengers and employees)" - This use case aimed to optimize existing public transport lines and potentially introduce new lines in the Zagreb airport catchment area to serve residents, passengers, and airport employees.
2. "Strategic planning of the airport gravitational areas and catchment zones" - This use case was developed to identify real catchment areas, enhance the sustainability of transport connections in those areas, and analyse potential areas not currently served by the airport but expected to be.
3. "Airline strategic planning" - The goal of this use case was to gather insights on travel demand for the planning of airline operations based on information about the initial country of origin and destination of airline users.
4. "Analytics of migration and retention habits of passengers" - This use case was developed to provide an analysis of migration patterns and retention habits of passengers, allowing for the distinction between passenger types (tourist, business users) to optimize services/offering for targeted users.
5. "Transport demand prediction" - This use case aimed to enable the analysis of transport demand for future predictions based on information on airport arrivals and departures, as well as the utilization of public transport services serving the airport area, including transport-on-demand providers. Historical data on migration patterns of airport passengers following aircraft arrivals (e.g. the number of passengers leaving the airport with a provider of on-demand transport or public transport) would be used to predict demand for specific transport modes/services (e.g. on-demand transport, rent-a-car, etc.).

The analysis results have offered a clear understanding of the current situation in the analysed area, enabling stakeholders to identify several shortcomings in the existing transport offerings, especially in the City of Zagreb, and to pinpoint potential improvements. In the future, after implementing measures to address the

identified shortcomings, the same tool and analytical use cases can be utilized to analyse the impact of those actions.

The identified analytical use cases were integrated as part of the newly developed IT Solution (platform) for traffic flow optimization, and the results were demonstrated for the Zagreb Airport area for all use cases and for the Paris Airport area for a limited number of use cases. This tool is expected to demonstrate its general applicability since it considers local data sources and specificities. Thanks to its modularity and flexibility, the tool can support additional use cases that rely on similar data sets with the primary goal of facilitating decision-making based on objective measurements.

Ultimately, the project has developed processes capable of calculating Key Performance Indicators as defined in the OLGA Project. The IT solution can calculate four KPIs (MS1.1, MS2.2, SOC3, GHG) that are related to the defined use cases.

In conclusion, this project has laid the groundwork for establishing a digital foundation for modern, data-driven, and multimodal mobility tools across Europe, drawing from harmonized traditional and innovative mobility data sources. Given the vital role of the mobility sector in European strategies, digital tools are crucial for ensuring a safer, more sustainable, intelligent, and resilient transport system. After three years of research and collaboration, it is evident that there are still challenges to overcome in order to ensure the full, harmonized implementation of the newly developed IT Solution. The long-term continuation of the newly developed IT solution will involve a focus on further developing standardized processes for collecting the status and format of data and metadata, and where they do not exist, proposing and implementing actions to promote non-conventional data, whose usability and value have been indisputably confirmed. Additional interaction and cooperation with external sectoral stakeholders will be necessary to define coordinated and agreed data accessibility and usability strategies, as well as strategic connections to support analytical use cases. The proposed IT solution, with both existing and new and improved analytical use cases, should serve as the foundation for data-driven decision-making.

6.1 Technology Readiness Level assessment

ENT was leading the development of an IT solution designed for multimodal traffic optimisation, followed by large-scale validation and replication. The innovation itself is mainly represented by the integration of analytics and fusion modules into a single platform, using data from different locally available sources, while ensuring data reliability, security and continuous improvement of transport management through ML and AI algorithms.

During the project implementation, Technology Readiness Level TRL 7 - Demonstration in operational environment was achieved.

IT platform prototype of the system was demonstrated as fully functional in an operational level, where all its components were integrated and tested. Goal of this stage was validation of the performance of the technology and all the operational/functional requirements were validated.

Following activities were achieved:

TRL 6

- System setup according to defined architecture – Fulfilled.
- Collection, utilisation and fusion of planned data sources (ZAG) – Fulfilled.
- Collection, utilisation and fusion of available data sources (CDG) – Fulfilled
- System demonstrated in a relevant environment (MS2.2. – Data from three different sources successfully fused in the platform – Demonstration in the workshop) – Fulfilled.

TRL7

- Environment for system implementation in operational environment prepared (Oracle cloud) – Fulfilled.
- System migration in operational environment at MZLZ – Fulfilled.
- Collection, utilisation and fusion of planned and available data sources in operational environment (ZAG, CDG) – Fulfilled
- Exposure of processed data through API interface - API Interface prepared for exposure– Fulfilled

7 References

[1] M. Tica, I. Štimac, K. Vidović, S. Vojvodić, and I. Stipanović, "Analytics Use Cases for Landside Traffic Optimization in the Catchment Area of the Airport: Case Study of Zagreb Airport," in *2023 46th International Convention on Information, Communication and Electronic Technology (MIPRO)*, Opatija, Hrvatska, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10159821/

[2] D. Medved, D. Blažinić, V. Galijan, and N. Antolović, "Evolution of data sources for integrated data-driven urban mobility management," *Transp. Res. Procedia*, vol. 64, pp. 68–75, 2022, doi: 10.1016/j.trpro.2022.09.009.

[3] R. Radović, I. Marasović, V. Čačković, D. Pleština, D. Keresteny, and Z. Anić, "The concept of a data aggregation platform in the function of a decision-making system for urban mobility management," in *Transportation Research Procedia*, 2022, pp. 53–59. doi: 10.1016/j.trpro.2022.09.007.

[4] M. Draganić *et al.*, "Requirements on applications within the decision-making system for urban mobility management," *Transp. Res. Procedia*, vol. 64, pp. 60–67, 2022, doi: 10.1016/j.trpro.2022.09.008.

[5] A. Galloni, B. Horváth, and T. Horváth, "Real-time Monitoring of Hungarian Highway Traffic from Cell Phone Network Data," in *ITAT 2018 Proceedings*, 2018, pp. 108–115.

[6] G. Kos, P. Brlek, and K. Vidovic, "SHARED SPACE CONCEPT IN LOCAL COMMUNITIES: CASE STUDY," in *8TH INTERNATIONAL CONFERENCE ON ROAD SAFETY IN LOCAL COMMUNITY*, Valjevo, 2013, pp. 1–8.

[7] J. Goulding, "Best Practices and Methodology for OD Matrix Creation from CDR Data," *NLAB, Univ. Nottingham*, vol. 44, no. 0, pp. 1–37, 2018.

[8] B. Furletti *et al.*, "Use of mobile phone data to estimate mobility flows. Measuring urban population and inter-city mobility using big data in an integrated approach," *Sis*, pp. 1–10, 2014, [Online]. Available: http://www.sistan.it/fileadmin/Repository/Home/IMMAGINI/01_In_evidenzaSIS_Cagliari_2014.pdf

[9] F. Yaghoubi, A. Catovic, A. Gusmao, J. Pieczkowski, and P. Boros, *Traffic Flow Estimation using Machine Learning and 4G/5G Radio Frequency Counters*, vol. 2022-June, no. 1. Association for Computing Machinery, 2022. doi: 10.1109/VTC2022-Spring54318.2022.9860858.

[10] M. Fahmideh and D. Zowghi, "An exploration of IoT platform development," *Inf. Syst.*, vol. 87, no. ثق ثقثقثق, p. 2020, ثقثقثقثق, doi: 10.1016/j.is.2019.06.005.

[11] I. Ganchev, Z. Ji, and M. O'Droma, "A generic IoT architecture for smart cities," *IET Conf. Publ.*, vol. 2014, no. CP639, pp. 196–199, 2014, doi: 10.1049/cp.2014.0684.

[12] E. Patti and A. Acquaviva, "IoT platform for Smart Cities: Requirements and implementation case studies," in *2016 IEEE 2nd International Forum on Research and Technologies for Society and Industry Leveraging a Better*

Tomorrow, RTSI 2016, 2016. doi: 10.1109/RTSI.2016.7740618.

[13] R. Roshan, A. Sharma, and O. P. Rishi, "IoT Platform for Smart City: A Global Survey," *Adv. Intell. Syst. Comput.*, vol. 841, pp. 197–202, 2019, doi: 10.1007/978-981-13-2285-3_24.

[14] R. Schwartz, M. Naaman, and Z. Matni, "Making sense of cities using social media: Requirements for hyper-local data aggregation tools," *AAAI Work. - Tech. Rep.*, vol. WS-13-04, pp. 15–22, 2013.

[15] S. Henning and W. Hasselbring, "Scalable and Reliable Multi-dimensional Sensor Data Aggregation in Data Streaming Architectures," *Data-Enabled Discov. Appl.*, vol. 4, no. 1, 2020, doi: 10.1007/s41688-020-00041-3.

[16] A. Nechifor, M. Albu, R. Hair, and V. Terzija, "A flexible platform for synchronized measurements, data aggregation and information retrieval," *Electr. Power Syst. Res.*, vol. 120, pp. 20–31, 2015, doi: 10.1016/j.epsr.2014.11.008.

[17] J. Poncela *et al.*, "Smart cities via data aggregation," *Wirel. Pers. Commun.*, vol. 76, no. 2, pp. 149–168, 2014, doi: 10.1007/s11277-014-1683-5.

[18] B. Pourghebleh and N. J. Navimipour, "Data aggregation mechanisms in the Internet of things: A systematic review of the literature and recommendations for future research," *J. Netw. Comput. Appl.*, vol. 97, pp. 23–34, 2017, doi: 10.1016/j.jnca.2017.08.006.

[19] R. Radović, D. Pleština, and V. Čačković, "Izvještaj o provedenom istraživanju izlaznog pokazatelja aktivnosti 5.1.1 - Uspostava i definiranje tehnološkog koncepta platforme za agregaciju podataka u funkciji odlučivanja u gradskom i multimodalnom prometu te urbanoj mobilnosti," Zagreb, 2021.

[20] V. L. Pérez and R. Schüler, "The Delphi Method as a tool for information requirements specification," *Inf. Manag.*, vol. 5, no. 3, pp. 157–167, 1982, doi: 10.1016/0378-7206(82)90022-2.

[21] R. Bossom and P. Jesty, "Extend FRAMEwork architecture for cooperative systems 15 – FRAME Architecture – Part 5: FRAME Architecture Methodology," 2011.

[22] P. H. Jesty and R. A. P. Bossom, "Using the FRAME Architecture for planning integrated Intelligent Transport Systems," *2011 IEEE Forum Integr. Sustain. Transp. Syst. FISTS 2011*, pp. 370–375, 2011, doi: 10.1109/FISTS.2011.5973610.

[23] K. Vidovic, M. Sostaric, A. Blavicki, and F. Sirovica, "Validation Points in Process of Urban Mobility Assessment Using Telecom Big Data Analytics," in *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, IEEE, Sep. 2021, pp. 1086–1090. doi: 10.23919/MIPRO52101.2021.9596920.