MOMENTUM

Deliverable D3.1

Data Inventory and Data Quality Assessment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 815069.

Table of Contents

Table Index

Figure Index

Summary sheet

Deliverable No.	3.1
Project Acronym	Momentum
Full Title	Modelling Emerging Solutions for Urban Mobility
Grant Agreement No.	815069
Responsible Author(s)	Antonio David Masegosa (Deusto) Pablo Fernández (Deusto) Itziar Salaberria (Deusto) Asier Moreno (Deusto) Javier Burrieza (Nommon) Josep María Salanova Grau (CERTH) Neofytos Boufidis (CERTH) Cristina Valdés (EMT Madrid) Eli Nomes (Leuven City) Christian Heil (Regensburg City) Marco Krakowitzer (SMO) Rodric Frederix (TML) Péter Pápics (TML) SanthanaKrishnan Narayanan (TUM) Constantinos Antoniou (TUM) Tamara Djukic (Aimsun)
Peer Review	Sergio Fernández (EMT)
Quality Assurance Committee Review	General Assembly
Date	24/03/2020
Status	Final draft
Dissemination level	Public
Abstract	The objective of the present document is to review the available data sources to characterise mobility in each of the four cases studies considered in this project (Madrid, Regensburg, Leuven and Thessaloniki), assess their strengths and weaknesses, and determine their potential usability for MOMENTUM. The outcomes of this analysis are a data inventory with more than 80 data sources available for MOMENTUM in each category and sub-category; a data quality assessment of each identified data source in terms of reliability, sample size, geographical and temporal scope, geographical and temporal granularity,

	completeness, validity and accessibility; and an analysis of the potential uses that each of this data sources may have for MOMENTUM according to its characteristics.
Version	Issue 1 Draft 5
Work Package No.	3
Work Package Title	Data Collection and Analysis
Programme	Horizon 2020
Coordinator	EMT Madrid
Website	www.h2020-momentum.eu
Starting date	01/05/2019
Number of months	36

Project partners

Organisation	Country	Abbreviation
EMPRESA MUNICIPAL DE TRANSPORTE DE MADRID SA	Spain	EMT
NOMMON SOLUTIONS AND TECHNOLOGIES SL	Spain	NOMMON
DIMOS THESSALONIKIS	Greece	THESS
ETHNIKO KENTRO EREVNAS KAI TECHNOLOGIKIS ANAPTYXIS	Greece	CERTH
STAD LEUVEN	Belgium	LEUVEN
TRANSPORT & MOBILITY LEUVEN NV	Belgium	TML
STADT REGENSBURG	Germany	REGENSBURG
TECHNISCHE UNIVERSITAET MUENCHEN	Germany	TUM
AIMSUN SL	Spain	AIMSUN SL
POLIS – PROMOTION OF OPERATIONAL LINKS WITH INTEGRATED SERVICES, ASOCIATION INTERNATIONALE	Belgium	POLIS
UNION INTERNATIONALE DES TRANSPORTS PUBLICS	Belgium	UITP
UNIVERSIDAD DE LA IGLESIA DE DESUTO – ENTIDAD RELIGIOSA	Spain	UDEUSTO

Document history

Version	Date	Organisation	Main area of changes	Comments
Issue 1 Draft 1	20/02/2020	Deusto	Initial version	
Issue 1 Draft 2	26/02/2020	Deusto	All sections	Second draft with new sections and corrections/comments from EMT and TUM
Issue 1 Draft 3	05/03/2020	Deusto	All sections	Consortium internal review
Issue 1 Draft 4	06/03/2020	Deusto	Summary sheet, Sections 2.2, 4.3.2, 5.9.2, and 5.10.2 and Appendix B	Name spelling and small corrections
Issue 1 Draft 5	24/03/2020	Deusto	Appendix B	Small corrections. Appendix A has been removed to make the document public, as it may contain sensitive information.

List of acronyms

GTFS	General Transit Feed Specification
MaaS	Mobility-as-a-Service
JSON	JavaScript Object Notation
SHP	Shape file
WP	Work Package
PT	Public Transport
PDF	Portable Document Fomat
SMO	das Stadtwerk Regensburg.Mobilität GmbH
OSM	Open Street Maps
NDA	Non-Disclosure Agreement
CSV	Comma Separated Values

Executive Summary

The overall goal of the MOMENTUM project is to develop a set of mobility data analysis and exploitation methods, transport models and planning and decision support tools able to capture the impact of new transport options and ICT-driven behavioural changes on urban mobility, in order to support local authorities in the task of designing the right policy mix to exploit the full potential of emerging mobility solutions.

The objective of the present document is **to review the available data sources to characterise mobility** in each of the four cases studies considered in this project (Madrid, Regensburg, Leuven and Thessaloniki), **assess their strengths and weaknesses**, and determine their **potential usability for MOMENTUM**. To structure and order the collection and analysis of the data sources, we defined five categories: **Transport Supply, Transport Demand, Maps & Cartography, Socio-Demographic and Travel Times**. Below we list the data source sub-categories that were reviewed for each of these categories:

- **Transport supply**: Public Transport Schedules and Lines, Transport Network, Taxi Service, Car/Moto-Sharing, Bike-sharing, MaaS and Parking
- **Transport demand**: Public Transport Smart Card, Other Public Transport Demand Data, Bike-sharing, Cycling, Pedestrian, Mobility Surveys, Telecom, MaaS, Car/Moto Sharing, Traffic, Taxi Service, Social Media and Parking.
- **Maps & Cartography**: Land Use, Weather, Social, Cultural or Sportive Events, Points of Interest.
- **Socio-demographic**: Demographic Statistics, Income statistics, Tourism statistics, Car Ownership statistics, Labor market statistics, House price statistics and other socio-demographic statistics.
- **Travel times**: Travel Time Data.

These data sources have been described, and the **information available has been analysed by filling a factsheet** with the following information:

- **General information:** Identification of the database and how to access it, Database name, Link to data.
- **Description:** Brief description of the dataset, reliability, potential use in MOMENTUM, actual use of the data, etc.
- **Availability:** Relevant information about owner and readiness to use. Owner, Access conditions, Data access, Availability within the project, Terms of use, etc.
- **Data resolution and use within MOMENTUM:** Temporal and geographical characteristics of the information provided. Data format, Temporal granularity, Temporal scope, Geographical granularity, Geographical scope, data aggregation, etc.
- **Comments:** Quality issues identified and other relevant information.
- **Information in the data source:** Descriptive list of the information contained in the data source.

The outcomes of this analysis are:

- A **data inventory with more than 80 data sources** available in total for MOMENTUM. They have been classified into categories and sub-categories.
- A **data quality assessment** of each identified data source in terms of reliability, sample size, geographical and temporal scope, geographical and temporal granularity, completeness, validity and accessibility.
- An **analysis of the potential uses** that each of these data sources may have for MOMENTUM according to its characteristics.

1 Introduction

1.1 Scope and Objective

The overall goal of the MOMENTUM project is to develop a set of mobility data analysis and exploitation methods, transport models and planning and decision support tools able to capture the impact of new transport options and ICT-driven behavioural changes on urban mobility, in order to support local authorities in the task of designing the right policy mix to exploit the full potential of emerging mobility solutions.

In order to achieve this general goal, a number of particular objectives were set that encompasses the development of a conceptual framework for assessing the impacts of new mobility options; the collection and analysis of heterogeneous data sources to characterise mobility; the formulation, calibration and validation of new models capable of capturing the effects of new mobility options; and the development of decision support tools that integrate these data and modelling improvements in the urban policy cycle.

This document falls under the second of the particular objectives listed above. It aims to provide an inventory of the different data sources available to the project and analysing their characteristics. In this way, this deliverable collects the results of the work done in the first part of WP3 "Data Collection and Analysis", and in particular to the task T3.1 Data collection and quality assessment. In this way, the specific objectives of this document are:

- **Review the available data sources to characterise mobility** in each of the four cases studies considered in this project: Madrid, Regensburg, Leuven and Thessaloniki. To this end, five different categories of data sources were considered: Transport Supply, Transport Demand, Maps & Cartography, Socio-Demographic and Travel time.
- **Assess the strengths and weaknesses of these data sources** in terms of reliability, sample size, geographical and temporal scope, geographical and temporal granularity, completeness, validity and accessibility.
- Determine their **potential usability for MOMENTUM** of each data source according to its characteristics.

1.2 Structure of the Document

The document is structured as follows:

- **Section 2** provides an overview of the data sources that have been reviewed, organised according to the different categories and sub-categories identified.
- **Section 3** presents the methodology followed for data quality assessment.
- **Section 4** reviews the results of the quality assessment of the Transport Supply data sources
- **Section 5** presents the results of the quality assessment of the Transport Demand data sources
- **Section 6** presents the results of the quality assessment of the Maps & Cartography data sources
- **Section 7** presents the results of the quality assessment of the Socio-Demographic data sources
- **Section 8** presents the results of the quality assessment of the Travel Time data sources
- **Section 9** discusses the main conclusions gathered from this deliverable
- **Appendix A** presents completeness, validity and simple exploratory data analysis of most relevant data sources.

1.3 Reference and applicable documents

Applicable documents:

[I] Grant Agreement No 815069 MOMENTUM – Annex 1 Description of the Action.

[II] MOMENTUM Consortium Agreement, Issue 1, April 2019.

[III] MOMENTUM D1.1 Project Plan, June 2019

[IV] MOMENTUM D1.2 Data Management Plan and Open Data Policy, November 2019

2 Overview of Data Sources

2.1 Definition and Classification of Potential Data Sources to be collected

As stated above, one of the objectives of WP3 is the collection and analysis of heterogeneous data sources to characterise mobility. In a more specific way, WP3 aims at gathering, for each of the four case study cities, different datasets with the potential to provide new insights on urban travel behaviour, including passively collected data from mobile devices (e.g., mobile phone records), sensor data (traffic counts, parking data, etc.), data on the use on new transport services, and more conventional data, such as mobility surveys. From this data, the final objective is to extract information such as mobility needs and preferences of different population segments, how people relate to the spatial structure and the socio-economic organisation of the city, or the potential of new forms of transport to satisfy the identified mobility needs.

With this idea in mind, before starting data collection, we first defined a list of potential data sources to be collected to fulfil the objectives stated above. To structure and order the collection and analysis of the data sources, we defined five general categories: Transport Supply, Transport Demand, Maps & Cartography, Socio-Demographic and Travel Times. Below we list the sub-categories that were considered for each of these general categories:

Data Category	Data Sub-category
Transport Supply	Public Transport Schedules and Lines
	Transport Network
	Taxi Service
	MaaS[1]
	Car/Moto-Sharing
	Bike-sharing
	Parking

[1] Although this subcategory of data source was first considered for its relevance to MOMENTUM, it was not included in the subsequent analyses because there were no reliable data sources available within this category in any of the four case studies.

Data Category	Data Sub-category
Transport Demand	Public Transport Smart Card
	Other Public Transport Demand Data
	Bike-sharing
	Cycling
	Pedestrian
	Mobility Surveys
	Telecom Data
	MaaS[1]
	Car/Moto Sharing
	Traffic data
	Taxi Service
	Social Media
	Parking
Maps & Cartography	Land Use
	Weather
	Social, Cultural or Sportive Events
	Points of Interest

[1] Although this subcategory of data source was first considered for its relevance to MOMENTUM, it was not included in the subsequent analyses because there were no reliable data sources available within this category in any of the four case studies.

Data Category	Data Sub-category
Socio-Demographic	Demographic statistics
	Income statistics
	Tourism statistics
	Car Ownership
	Labour market statistics
	House price statistics
	Other socio-demographic statistics
Travel Time	Travel Time Data

2.2 Availability of Data Sources within the project

Below, a data source table summary for each city is presented. The column “Data Provider” refers to the institution that provides the data. In contrast, the column “Private/Public Institution” indicates whether the data provider is a public or a private institution.

2.2.1 Madrid

Data type	Data Scources	Type	Data Provider	Private/ Public Institution	Temporal Scope	Geographical Scope	Temporal Granularity	Geographical Granularity
Transport supply	Public Transport Schedules and Lines	Scheules and lines for Underground, , Urban buses, Interurban buses and Suburban trains	EMT, Madrid Regional Transport Consortium	Public	2019	City + Metrop. Area	Minutely	PT stop
		REST API for Urban Buses	EMT	Public	Real-time	City	Minutely	Bus stop
	Transport Network	Street Map	Regional government	Public	N/A	Madrid region	N/A	Street level
	Taxi Service Supply Data							
	Car/Moto-Sharing Data Supply							
	Bike-sharing Data Supply	Station-Based Bike-Sharing System (stations locations, bike and dock avail.)	EMT	Public	2018-2019	City centre	Hourly	Station level
	MaaS Data Supply							
	Parking Data Supply	Public off-street parking lots	EMT	Public	2016-2019	City	Minutely	Parking level
Transport demand	Public Transport Smart Card Data	Smart Card Validations for urban buses	EMT	Public	2006-2019	City	Validation time	Bus stop
	Other Public Transport Data							
	Bike-sharing Data Demand	Station-base bike-sharing	EMT	Public	2017-2019	City centre	Hourly	Individual level
	Cycling Data	Counts from fibre-optic sensors	City Council	Public	Jan - Jun 2019	City (19 locations)	15 mins	Sensor locations
	Pedestrian Data	Counts from Fiber-optic sensors	City Council	Public	Jan - Jun 2019	City (19 locations)	15 mins	Sensor locations
	Mobility Surveys							
	Telecom Data	Call Detail Records	Orange (processed by Nommon)	Private	2016-2019	City	20-30 mins	Mobile network antenna level
	MaaS Data Demand							
	Car/Moto-Sharing Data Demand							
	Traffic Data	Loop sensors	City Council	Public	2013-2019	City	15 mins	At sensor level
	Taxi Service Demand Data							
	Last Mile Logistics Data							
	Social Media Data							
	Parking Data Demand	Periodical snapshots of parking occupancy	EMT	Public	2016-2019	City	3 mins	At parking level

Maps / Cartography	Land Use Data	National Land Cover GIS Database	SIOSE, City Council	Public	2005, 2011, 2014	Madrid region	N/A	Land use polygons
	Weather Data	Weather stations	AEMET	Public	Apr - Dec 2019	Country	Hourly	Weather station level
	Social, Cultural or Sportive Events							
	Points of Interest							
Socio-demographic	Demographic statistics	Population Register	INE	Public	2008-2019	Country	Yearly	Census tract
	Income statistics	Income Tax statistics from National Tax Agency	INE	Public	2015-2017	Country	Yearly	Census tract
	Tourism statistics	Surveys	City Council	Public	N/A	City	N/A	N/A
	Car Ownership	Vehicle Tax Database	City Council	Public	2016-2018	City	Yearly	Census tract
	Labor Market Statistics	Income Tax statistics from National Tax Agency	INE	Public	2015-2017	Country	Yearly	Census tract
	House price statistics							
	Business statistics	Census of remises and Activities	City Council	Public	N/A	City	Yearly	Neigbourhood
	Other socio-demograhic data							
Travel times	Travel Time Data	Maps API	Google	Private	N/A	Madrid region	Wednesday at 7:30AM, 5:30AM and 12:AM	Centroid pair

2.2.2 Regensburg

Data type	Data Scources	Type	Data Provider	Private/ Public Institution	Temporal Scope	Geographical Scope	Temporal Granularity	Geographical Granularity
Transport supply	Public Transport Schedules and Lines	Time deviations from planned schedules	SMO	Public	2016-2019	City	Hourly	Bus stop
		Schedules and lines for urban buses	SMO	Public	2019	City	Minutely	Bus stop
	Transport Network	Extraction from Land Cover GIS	City Council	Public	N/A	City	N/A	Street level
	Taxi Service Supply Data							
	Car/Moto-Sharing Data Supply	Station-based Car-sharing	REWAG	Public	2017-2019	City+Metropolitan Area	Monthly	Station level
	Bike-sharing Data Supply							
	MaaS Data Supply							
	Parking Data Supply							
Transport demand	Public Transport Smart Card Data							
	Other Public Transport Data	Passenger counts from sensorized buses	SMO	Public	2017-2019	City	Hourly	Bus stop
	Bike-sharing Data Demand							
	Cycling Data	Manual Counts	City Council	Private	6 days in September 2018	Oldtown	Hourly	Counting locations
	Pedestrian Data	Manual Counts	City Council	Private	6 days in September 2018	Oldtown	Hourly	Counting locations
	Mobility Surveys	Household survey	City Council	Public	Feb 2018 – Jan 2019	City	N/A	District level
	Telecom Data							
	MaaS Data Demand							
	Car/Moto-Sharing Data Demand	Station-based car-sharing	REWAG	Public	2016 -2019	City+Metropolitan Area	Trip/Booking Event	Station level
	Traffic Data	Loop sensors	Regional Governmet	Public	2018-2019	Regensburg's ring	Hourly	Sensor location
		Floating car data	(e.g INRIX)	Private	One week	City	Minutely	Segment level
	Taxi Service Demand Data							
	Last Mile Logistics Data							
	Social Media Data							
	Parking Data Demand							

Maps / Cartography	Land Use Data	City GIS Database	City Council	Private	2017	City	N/A	Homogeneous land use polygon
	Weather Data	Weather info API	DarkSky.net	Private	From 2015	City	Hourly	Weather station
	Social, Cultural or Sportive Events							
	Points of Interest							
Socio-demographic	Demographic statistics	Population Register	City Council	Public	2017-2019	City	Yearly	City districts
	Income statistics	Average Household income statistics from Regional Government	City Council	Public	1995-2019	City	Yearly	City
	Tourism statistics	Surveys	City Council	Public	Since 2017	City	Yearly	City
	Car Ownership	Vehicle Tax Database	City Council	Public	1998-2019	City	Yearly	City districts
	Labor Market Statistics	Statistics	City Council	Public	Since 2017	City	Yearly	City districts
	House price statistics	Rental Index/Real State Report	City Council/Sparkasse Regensburg	Public	2009-2016	City	Yearly	City districts
	Business statistics	Census of remises and Activities	City Council	Public	1998-2019	City	Yearly	City
	Other socio-demograhic data	Statistics	City Council	Public	Sinde 2017	City	Yearly	City districts
Travel times	Travel Time Data							

2.2.3 Leuven

Data type	Data Scources	Type	Data Provider	Private/ Public Institution	Temporal Scope	Geographical Scope	Temporal Granularity	Geographical Granularity
Transport supply	Public Transport Schedules and Lines	Schedules and lines for buses	De Lijn	Public	2015-2019	Flanders	Daily	City
	Transport Network	Road Network Database	City Council	Public	N/A	Flanders	N/A	Road segment
	Taxi Service Supply Data							
	Car/Moto-Sharing Data Supply							
	Bike-sharing Data Supply	Station-Based Bike-Sharing System (station location, bike avail.)	Bluebikes	Private	2015-2019	City	Yearly	Bike station
	MaaS Data Supply							
	Parking Supply Data	Public off-street and shop&go parking lots	City Council	Public	2019	City	Yearly	Parking facility
Transport demand	Public Transport Smart Card Data							
	Other Public Transport Data							
	Bike-sharing Data Demand	Station-base bike-sharing	Bluebikes	Private	2015-2018	Only one bike haring point in train station	Weekly	Station level
	Cycling Data	Counts from fibre-optics sensors	City Council	Public	2015-2017	City (5 locations)	Hourly	Sensor locations
		Counts from Crowdsourced Camera sensors	Teelram	Public	2019	City	Hourly (daytime)	Sensor locations
	Pedestrian Data	Countings from Crowd- sourced Camera sensors	Teelram	Public	2019	City	Hourly (daytime)	Sensor locations
		Wifi detectors	City Council	Public	Since 2017	City Center	Hourly	Sensor locations
	Mobility Surveys	City Monitor	City Council	Public	May 2 – June 6 2017	City	N/A	8 sub-areas
		OVG	Flemish Government	Public	2017	Flanders	N/A	Postal code
		Student Survey	KU Leuven	Public	2017	City	N/A	Three sub-areas: Brussels, Flanders and Wallonia
		WWV	Belgian Government	Public	July 1 2017 - Jan 31 2018	City	N/A	Company premises
	Telecom Data							
	MaaS Data Demand							
	Car/Moto-Sharing Data Demand							
	Traffic Data	Countings from Crowd- sourced Camera sensors	Teelram	Public	2019 (daytime)	City	Seconds	Road segment
		Intersection counts	AWS	Public	2019	City	15 minutes	Turns
	Taxi Service Demand Data							
	Last Mile Logistics Data							
	Social Media Data							
	Parking Data Demand	Vehicle entry/exit events	City Council	Public	2018-April 2019	City	Entry/Exit event	Individual parking facilities

Maps / Cartography	Land Use Data	City GIS Database	City Council	Public	N/A	City	N/A	Building Block
	Weather Data	Weather sensors	Weatherunder ground	Private	From 2015	City	Hourly	Weather station
	Social, Cultural or Sportive Events	Crowd-sourced event database	Uitdatabank	Private	2006-2019	City	N/A	Event location
	Points of Interest	City GIS Database	City Council	Public	2019	City	N/A	Building Block or POI
Socio-demographic	Demographic statistics	Population Register	STATBEL	Public	2009-2019	City	Yearly	Statistical Sector
	Income statistics	Income tax data from National Tax Agency	STATBEL	Public	2005-2017	Country	Yearly	Statistical Sector
	Tourism statistics	Survey	STATBEL	Public	2014-2018	Country	Yearly	City
	Car Ownership	Survey	City Council	Public	2019	City	Yearly	Statistical Sector
	Labor Market Statistics	Statistics	KSZ	Public	2005-2019	Country	Yearly	Statistical Sector
	House price statistics	Real State Sales Statistics	STATBEL	Public	2016-2019	Country	Yearly	Statistical Sector
	Business statistics	Statistics	KBO	Public	2016-2019	Country	Yearly	Statistical Sector
	Other socio-demograhic data	Statistics	City Council	Public	2019	City	Yearly	Statistical Sector
Travel times	Travel Time Data	Travel times	City Council	Public	2019	City	Specific times during the day	Specific routes

2.2.4 Thessaloniki

Data type	Data Scources	Type	Data Provider	Private/ Public Institution	Temporal Scope	Geographical Scope	Temporal Granularity	Geographical Granularity
Transport supply	Public Transport Schedules and Lines	Schedules and lines for buses	CERTH-HIT	Public	2013-2018	City and suburbs	Minutely	PT stop
	Transport Network	Transport Network Database	CERTH-HIT	Public	N/A	City	N/A	Road Segment
	Taxi Service Supply Data	Taxi trips	CERTH-HIT/TaxiWAY	Public/ Private	2016-2019	City and suburbs	Seconds	Origin / dest. coords
	Car/Moto-Sharing Data Supply							
	Bike-sharing Data Supply	Station-Based Bike-Sharing System (stations locations, bike and dock avail.)	Brainbox Technologies S.A	Private	Since 2013	City	Minutely	Station level
	MaaS Data Supply							
	Parking Data Supply							
Transport demand	Public Transport Smart Card Data							
	Other Public Transport Data							
	Bike-sharing Data Demand	Station-base bike-sharing	Brainbox Technologies S.A	Private	Since 2013	City	Minutely	Station level
	Cycling Data							
	Pedestrian Data							
	Mobility Surveys	Household survey	CERTH-HIT	Public	2017-2018	Prefecture of Thessaloniki	N/A	Traffic zone
	Telecom Data							
	MaaS Data Demand							
	Car/Moto-Sharing Data Demand							
	Traffic Data	Floating car data (TaxiWAY)	CERTH-HIT/TaxiWAY	Public/ Private	2013-2019	City	15 minutes	Exact coordinates
	Taxi Service Demand Data	Taxi trips	CERTH-HIT/TaxiWAY	Public/ Private	2016-today	City and suburbs	Seconds	Exact coordinates
	Last Mile Logistics Data							
	Social Media Data	Geo-tagged check-in event	Facebook	Private	Since 2016	City	20 minutes	Venue location
	Parking Data Demand							

Maps / Cartography	Land Use Data	City GIS Database	CERTH-HIT & City Council	Public	N/A	City	N/A	Building block
	Weather Data	Weather sensors	CERTH-HIT	Public	2012-2019	City	Minutely	Sensors location
	Social, Cultural or Sportive Events							
	Points of Interest	City GIS Database	CERTH-HIT & City Council	Public	N/A	Perfecture of Thessaloniki	N/A	Exact coordinates
Socio-demographic	Demographic statistics	Population Census	CERTH-HIT	Public	1951-today	Country	Yearly	Building block
	Income statistics	Income statistics from Household Survey	CERTH-HIT	Public	2018	Prefecture of Thessaloniki	Yearly	Traffic Zone
	Tourism statistics							
	Car Ownership	Population Census	CERTH-HIT	Public	1951-2011	Country	Every 10 years	Building block
	Labor Market Statistics	Population Census	CERTH-HIT	Public	1951-today	Country	Every 10 years	Building block
	House price statistics							
	Business statistics							
	Other socio-demograhic data							
Travel times	Travel Time Data	Bluetooth sensors	CERTH-HIT	Public	2017-today	City	Seconds	Predefined paths

3 Quality Assessment Methodology

3.1 Description of the methodology

Once the data sources have been identified, it is necessary to analyse and classify all the information that is provided in each one of them. This analysis has been performed by filling, for each data source, the factsheet depicted in Figure 1.

The information recorded in the factsheet can be summarised as follows:

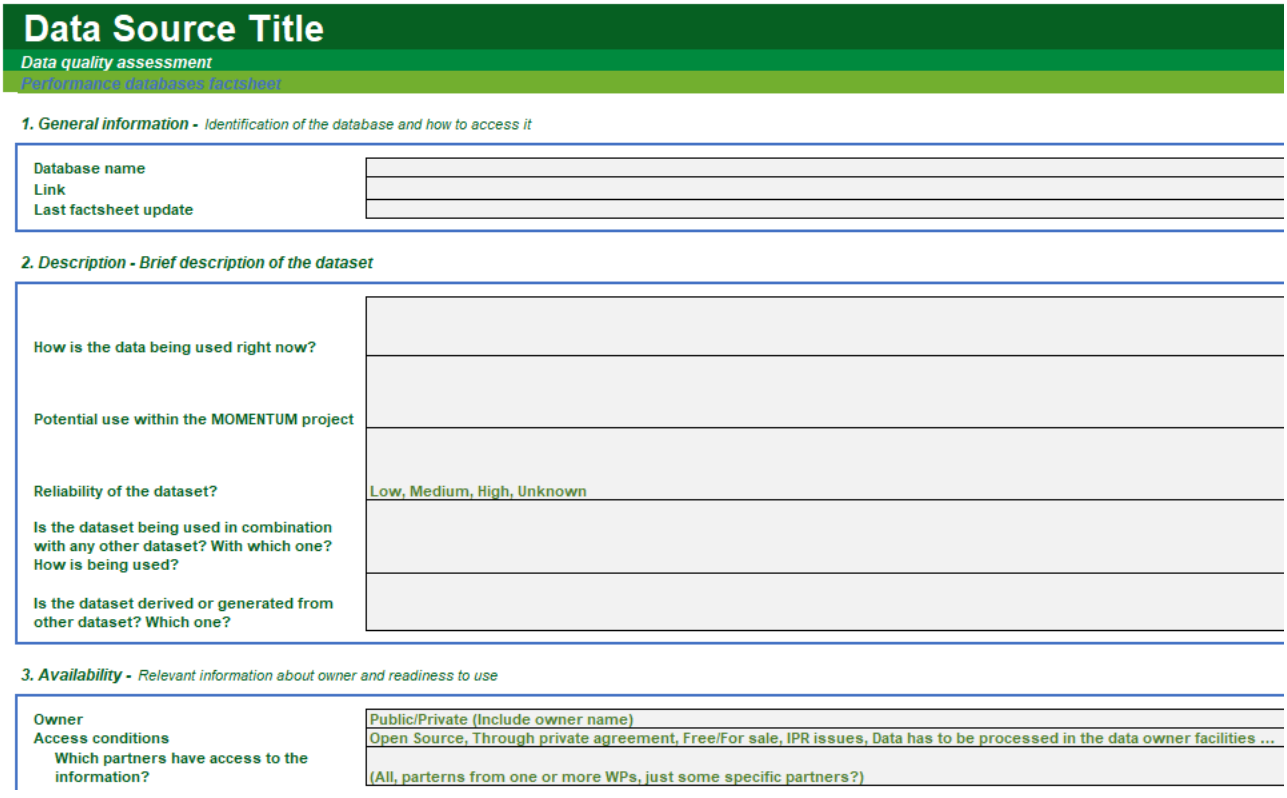
1. **General information**
    a. Database name
    b. Link
    c. Last factsheet update
2. **Description** - Brief description of the dataset
    a. How is the data being used right now?
    b. Potential use within the MOMENTUM project
    c. Reliability of the dataset?
    d. Is the dataset being used in combination with any other dataset? With which one? How is being used?
    e. Is the dataset derived or generated from other dataset? Which one?
3. **Availability** - Relevant information about owner and readiness to use
    a. Owner
    b. Access conditions
        i. Which partners have access to the information?
    c. Data access
        i. Data access technology
        ii. It is authentication required?
        iii. What type of authentication?
        iv. It is VPN required to access the data?
        v. Is there any data volume exchange limit?
        vi. Is there any API call limit per time?
        vii. Update frequency
    d. Availability within the project
    e. Will be the data available 100% of time?
    f. Terms of use
        i. Is the data sensitive and must it be anonymized or processed to comply with the GDPR?
        ii. It is already anonymized?
    g. Link to the data source
4. **Data resolution and use within momentum** - Temporal and geographical characteristics of the information provided. Fill only if these characteristics are common for the entire database.
    a. Data format
        i. Which is the format of the information?
        ii. Can the data format be chosen?
        iii. Is there documentation describing the fields of the data source?
        iv. Link to the data source documentation
        v. Is the data format proprietary? Is any licensed/purchased software necessary to view the data?
        vi. Is there any free / open source tool to view and transform the data?
        vii. Approximate size of the dataset

viii. Is any of the data related to any reference transport network?
ix. Is a closed or open reference transport network?

b. Temporal granularity
   i. Can temporal granularity be customized?
c. Temporal scope
   i. Can the temporal scope be customized?
d. Geographical granularity
   i. Can geographical granularity be customized?
e. Geographical scope
   i. Can the geographical scope be customized?
f. Is there any other filter applicable to the data source?
   i. Which filter/s?
g. Is there any aggregation applied to the datasets?
   i. Which kind of aggregation(s)?

5. **Comments** - Other relevant information
6. **Information in the data source** - Descriptive list of the information contained in the data source
7. **Data sample**

The completed factsheets for all the reviewed data sources are included in Appendix A.



Data Source Title

Data quality assessment

Performance databases factsheet

1. General information - *Identification of the database and how to access it*

Database name	
Link	
Last factsheet update	

2. Description - Brief description of the dataset

How is the data being used right now?	
Potential use within the MOMENTUM project	
Reliability of the dataset?	Low, Medium, High, Unknown
Is the dataset being used in combination with any other dataset? With which one? How is being used?	
Is the dataset derived or generated from other dataset? Which one?	

3. Availability - *Relevant information about owner and readiness to use*

Owner	Public/Private (Include owner name)
Access conditions	Open Source, Through private agreement, Free/For sale, IPR issues, Data has to be processed in the data owner facilities ...
Which partners have access to the information?	(All, parterns from one or more WPs, just some specific partners?)

Data access	
Data access technology	Web Service, Message Bus, File download from a remote server, Physical copy of a file, etc?
It is authentication required?	Yes / No
What type of authentication?	User + password, API Key, etc ?
It is VPN required to access the data?	Yes / No
Is there any data volume exchange limit?	Yes / No, how much volume?
Is there any API call limit per time?	Yes / No, which limit?
Update frequency	5 minutes, hourly, daily, etc.
Availability within the project	Available / In process / To explore. Private agreement signed, license(s) available, other..., and link or reference to the relevant document
Will be the data available 100% of time?	Yes / No. If not so, please describe the period during which it will be available
	Scope, period or events that might condition the use rights granted by the agreement, if any. Privacy and confidentiality requirements to b
Terms of use	anonymisation, etc)
Is the data sensitive and must it be anonymized or processed to comply with the GDPR?	Yes / No. If so, please indicate the data processing required to comply with GDPR and whether it will be addressed by the city or it must be
It is already anonymized?	Yes / No
Link to the data source	

4. Data resolution and use within momentum - *Temporal and geographical characteristics of the information provided. Fill only if these characteristics are common for the entire database.*

Data format	
Which is the format of the information?	json, xml, csv, kml, etc.
Can the data format be chosen?	Yes / No
Is there documentation describing the fields of the data source?	Yes / No
Link to the data source documentation	http://www.xxxxxx.xxx
Is the data format propietary? Is any licensed/purchased software neccesary to view the data?	Yes / No, which software is needed
Is there any free / open source tool to view and transform the data?	Yes / No, which one
Aproximate size of the dataset	Size in Mb, Gb
Is any of the data related with any reference transport network?	OSM, etc
Is a closed or open reference transport network?	Open / closed. If closed specity how this reference network can be provided
Temporal granularity	Minutely/Hourly/Daily/Weekly ...
Can temporal granularity be customized?	Yes / No

Temporal scope	Last 5 years, one month, this year, specific period of time, ...
Can temporal scope be customized?	Yes / No
Geographical granularity	Neighborhood, lane ...
Can geographical granularity be customized?	Yes / No
Geographical scope	Country, Region, City ...
Can geographical scope be customized?	Yes / No
Is there any other filter applicable to the data source?	Yes / No
Which filter/s?	
Is there any aggregation applied to the datasets?	Yes / No
Which kind of aggregation(s)?	

5. Comments - *Other relevant information*

6. Information in the data source - *Descriptive list of the information contained in the data source*

7. Data sample

Please provide a link to a sample data file hosted in any of youw known servers or upload new sample data files to the Google Drive Folder. The naming conve
DATASOURCEID.extension
If there are more than one file for each data source please use the following naming convention:
DATASOURCEID_001.extension

Figure 1: Data Factsheet template

Type & Data quality dimension		Case Study
Type		Specific typology of the data source
Reliability		High, Medium, Low
Temporal & Geographical Scope	Last update	Date, year or month
	Temp scope	Date range
	Geo scope	City, City+Metropolitan area, Region, Country, etc.
Granularity & Level of Detail	Temp gran	Yearly, Monthly, Daily, Minutely, Event, etc.
	Geo gran	Street, Building Block, Zone, District, City, etc.
	Level of Detail	Anonymized user ID, trip, aggregated measure, etc.
Availity & Accessibility		- Availability: Whole project, from a specific date, up to a specific date - Data format: CSV, Excel, JSON, SHP, etc. - Accessibility: Public, All partner, Specific partners, etc.
Relevance/Usability		High, Medium, Low

Table 1. Scheme of the summary table use of the general overview of the data quality assessment for each data source sub-category

3.2 Description of the general overview of the data quality assessment

To facilitate the understanding and visualization of the data quality assessment, for each of the sub-categories defined in Section 2.1, we have created a table summarizing the specific type of data source, the main characteristics that determine the quality of the data source, and the relevance and usability of the data source for MOMENTUM. The structure of this summary table is shown in Table 1.

As can be seen, the first column indicates the type and each of the data quality dimensions that we have considered most relevant. We describe each of these elements below:

- **Type:** Indicates the specific type of data source within the general category and subcategory considered.
- **Reliability:** Indicates the extent to which we can trust the data, from the point of view of its accuracy and precision.
- **Temporal & Geographical Scope:** Three sub-dimensions are considered here
  - Last update: the date when the data source was last updated
  - Temporal scope: Range of dates for which historical data is available
  - Geographical scope: a geographic area for which data would be available
- **Granularity and Level of Detail:** Three sub-dimensions are also considered here
  - Temporal granularity: indicates the minimum aggregation period in which the data would be available
  - Geographical granularity: shows the minimum geographical area for which data are aggregated
  - Level of Detail: given the temporal and geographic granularity of the data source, this dimension indicates whether there is only aggregated data for each granule (e.g. number of trips from a bike-sharing station at hour interval) or whether the data source allows for a higher level of detail (e.g. detail trip information with the anonymized user ids from a bike-sharing station at hour interval)
- **Availability and Accessibility:** in this case, there is a single cell containing three items:
  - Availability: indicates if the data source will be available during the whole project, from a specific date, until a particular time, etc.
  - Data format: indicates in which data format the data source is available
  - Accessibility: means if the data source is public, or in case it is not, which partners will have access to that data source and under which conditions.
- **Relevance/Usability:** it is a dimension that actually summarizes all the previous ones and even considers some more and defines the degree of usability of the data source within the project, given its characteristics in terms of data quality.

Finally, to improve and facilitate the visualization of those aspects that have a higher impact on the quality of a data source a colour code has been used in each cell, which indicates the impact that the value of that dimension for the data sour at hand has on the quality. Specifically, four levels are considered:

- **Green:** means that the degree of quality of the data source with respect to that dimension is good
- **Yellow:** indicates that the degree of quality of the data source with respect to that dimension is medium
- **Orange:** implies that the degree of quality of the data source with respect to that dimension is low and has a moderate impact on the usability of the data source
- **Red:** indicates that the degree of quality of the data source with respect to that dimension is deficient and has a medium impact on the usability of the data source.

4 Data Quality Assessment for Transport Supply Data Sources

4.1 Public Transport Schedules and Lines

4.1.1 Introduction

In the case study of **Madrid**, there are two data sources available with PT schedules and lines information. On the one hand, we have the data catalogue published on the website of the Madrid Transport Consortium[1], where GTFS[2] files with timetables, lines, routes, stops, pricing, etc. are publicly available for each of the PT services of the operators that make up this consortium, specifically Metro Madrid, EMT, Interurban Buses and Renfe suburban trains. These files are updated between one and two times a year and cover the PT of the city and region of Madrid. The last update took place in April 2019, and they can be accessed by downloading the ZIP file that in turn contains the different files that make up the GTFS standard. And on the other hand, we have a REST API provided by the EMT that offers, in real-time, also with data about EMT buses schedules and lines, but enriched with information about delays and incidents that affect the regular operation of this service.

For the city of **Leuven**, the information on the bus schedules and lines for the PT services operated by De Lijn (the PT company of the Flanders region in Belgium) are publicly available on the Open Mobility Data portal (previous TransitFeeds)[3] in GTFS format and covers the whole territory of Flanders. In this case, the data source is updated several times a month. Apart from this, Leuven City Council has available GTFS files from 2015 until now that are accessible for all MOMENTUM partners.

In **Regensburg**, the data provider is SMO (Stadtwerk Regensburg Mobilität GmbH), the PT operator of this city. In this case, the files containing the PT schedules and lines are also in GTFS format, although not publicly available as the previous ones. However, SMO has provided to MOMENTUM partners for its use for the whole project. The last update of this file took place in December 2019, although SMO informed us that the frequency of updating is not fixed in advance. In this case, the geographical scope is the city of Regensburg.

Apart from the above data source, for this case study, there is another one that includes aggregated information about time deviations of buses with respect to the planned schedule, for each bus line(route) and trip. It is currently used for planning purposes, and it is extracted by combining information from the buses GPS trajectory records and the GTFS files. This dataset is provided as a set of Excel files (one per bus route and aggregation period up to one hour. SMO keep records of this data source from 2016. The information available for each bus trip and aggregation period is the following (there is one row per each 20 meters travelled by bus): bus stop (if there is one at that point in the bus trip), distance in meters travelled by the bus from the starting point of the trip, mean time deviation with respect to the planned schedule in seconds (a negative value indicates a delay) and a value indicating how many trips were evaluated to obtain these statistics. The geographical scope, in this case, is also the city of Regensburg.

Finally, for the **case study in Thessaloniki**, the dataset available consists of the bus stop points location and the number of public transport lines that serve the Regional Unit of Thessaloniki. The raw public transport data used to produce the dataset are publicly available in the webpage[4] of the public transport operator in the region of Thessaloniki, namely Organization of Urban Transportation of Thessaloniki (OUTH), who is the owner of the data.

[1] https://datos.crtm.es/search?collection=Dataset
[2] https://developers.google.com/transit/gtfs
[3] http://transitfeeds.com/feeds
[4] http://oseth.com.gr/en/downloads/

It is currently used to feed the mobility planner at Thessaloniki's Urban Mobility Centre and other relevant, informative modules. Additionally, the dataset has been used to encode the operation of the public transport network, in transport simulation software (Sumo, Visum, Vissim, Aimsun), in multiple applications for mobility planning and assessment in the Metropolitan Area of Thessaloniki. The dataset has not been updated since 2018 because of internal problems in the organization. The dataset is currently available for CERTH and optionally to other specific partners under an NDA.

4.1.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid		Regensburg		Leuven	Thessaloniki
Type		Schedules and lines for Underground, Urban buses, Interurban buses and Suburban trains	REST API for Urban Buses	Time deviations from planned schedules	Schedules and lines for urban buses	Schedules and lines for buses	Schedules and lines for buses
Reliability		High	High	High	High	High	Medium/ High
Temp. & Geo Scope	Last update	December 2019	December 2019	February 2019	December 2019	December 2019	2018
	Temp scope	2019	Real-time	2016-2019	2019	2015 – 2019	2013 – 2018
	Geo scope	Madrid Region	City	City	City	Flanders	City + Metropolitan Area
Gran. & Level of Detail	Temp gran	Minutely	Minutely	Hourly	Minutely	Minutely	Minutely
	Geo gran	PT stop	Bus stop	Relative bus route location	Bus stop	Bus stop	Bus stop
	Level of Detail	N/A	N/A	N/A	N/A	N/A	N/A

Type & Data quality dim.	Madrid		Regensburg		Leuven	Thessaloniki
Availability & Accessibility	Whole project GTFS Public	Whole project JSON Public	Whole project Excel All partners	Whole project GTFS All partners	Whole project GTFS All partners	Whole project GTFS Spec. partners (NDA)
Relevance/ Usability	High	High	High	High	High	High

Table 2. General overview of the Data Quality Assessment of Public Transport Schedules and Lines

4.1.3 Identified issues

For the **Madrid Case Study,** the only relevant issue found is related to the GTFS file of the interurban buses because the schedules contained are not accurate. Considering the REST API available from EMT with urban buses schedules, the only problem it shows is that it does not follow a standard, such as RT-GTFS. In this way, its use would entail the design of ad-hoc tools for its use or for its transformation to GTFS format.

Regarding the data source for public transport schedules and lines in **Regensburg**, no relevant limitation was detected but some minor problems. On the one hand, the updates of the GTFS file are not done with a pre-established periodicity, which does not guarantee to have a relatively updated version of it. Anyway, in our case, the last update was done in December 2019, so it does not pose a problem for MOMENTUM. On the other hand, the geographical scope is only the city of Regensburg, with no other public transport data available out of this area.

As for the data source on time deviations in Regensburg, the main problem it presents is that the software containing these data allows only for manual extractions of aggregated data for one period of analysis at a time. If a time granularity of one hour is required, a manual extraction should be made for each hour of the period to be studied (e.g. a week of that would require 7x24 = 168 manual extractions). This issue makes obtaining data for long periods a laborious task.

The data source available for **Leuven** and **Thessaloniki** does not present any issue.

4.1.4 Relevance/Usability

Data sources regarding timetables and public transport lines for **Madrid** (except intercity buses), **Regensburg**, **Leuven** and **Thessaloniki** are of high relevance to MOMENTUM. In all cases, we have GTFS files available, which facilitates the use of off-the-self free tools for their processing. Furthermore, they have an excellent temporal and geographical scope. The potential uses of these data sources within the MOMENTUM project are, on the one hand, the modelling of PT supply in the respective case studies; and on the other hand, as a complementary or necessary data source for the analysis of the public transport demand in Madrid and Regensburg, the only two cities with PT demand data. Concretely, for this purpose, they must be combined with the data sources described in Sections 5.1 and 5.2., respectively.

As for the data source on time deviations available for **Regensburg**, its relevance to the project is high. It provides very accurate and detailed information about the time reliability of bus transport in Regensburg, although at an aggregated level (at least hourly). Its potential use in MOMENTUM is the study of the

influence of the variability of public transport in demand by the joint analysis of this data source with the one described in Section 5.2.

4.2 Transport Network

4.2.1 Introduction

Before starting with the particularities of the data sources of each case study, we would like to mention that in this category there is an open data source that would be universal to all the cities, which is Open Street Maps (OSM)[1]. OSM is a free, editable world map generated collaboratively and voluntarily, where crowdsourced content is available under an Open Database License. It is a map based on a topological data structure that models the transport network as a graph and allows to give very detailed information about it (e.g. intersections, street/road segments, lines, maximum speed, traffic light locations, etc.). It currently has more than 6 million registered users and 1.4 million contributors who perform around 16,000 weekly updates.

For the case study in **Madrid**, the additional transport network data source available is the street map of the Region of Madrid, which is publicly accessible on the website of this region's Statistics Institute[2]. It provides only the geometries of the road network, the road category and the name. It does not present a topological data structure, and it is updated annually (last version from 2018). The data of the transport network can be obtained in SHP format.

In the case of **Regensburg**, the available transport network has similar characteristics to those described for Madrid. This is an extraction of the road network layer from the governmental planning tool for land use. Since it is land use information, only the geometry, name and category of the road can be obtained. Like the previous one, it does not have a topological data structure. This map of the transport network was last updated in 2017, and its geographical scope is the city of Regensburg. It is a private data source that will be available to all partners in SHP format.

For **Leuven**, a transport network based on a topological data structure created by the Flanders Road and Transport Agency is available. The data source is called the Wegen Register (Road register) and contains all public transport ways including roads, bicycle lanes and footpaths in the region of Flanders. For each road segment, it offers information such as status, road category, number of lanes, road width, type of surface, etc. The accuracy of the data source according to the available documentation is medium (Leuven reported that it is similar to that of OSM) and it is publicly accessible in SHP format on the Government of Flanders website[3]. The last update is from September 2019.

Finally, in the case study of **Thessaloniki**, the available data on the transport network have been provided by CERTH-HIT, who owns and maintains digitized network files for the Metropolitan Area of Thessaloniki. Transport network data include the road network for the Regional unit of Thessaloniki, although a different level of detail (up to local roads) applies for the metropolitan area of Thessaloniki. The road network characteristics have been updated recently (2018) for the Municipality of Thessaloniki (as a part of the SUMP process). Public transport network has also been updated recently (2018). It is a private data source that will be available only to some specific partners under an NDA. The data about the transport network are available at a high level of detail and refer to the geometrical characteristics of the network (length, position in the network); traffic characteristics (number of lanes, free-flow speed, capacity per lane, transport modes allowed to use each road section, existence of dedicated bus lanes, traffic light programs etc.); and other descriptive characteristics such as the name of the road section.

[1] https://www.openstreetmap.org/
[2] http://www.madrid.org/nomecalles/DescargaBDTCorte.icm
[3] https://overheid.vlaanderen.be/informatie-vlaanderen/producten-diensten/wegenregister

4.2.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Street Map	Extraction from Land Cover GIS	Road Network Database	Transport Network Database
Reliability		High	High	Medium	High
Temporal & Geographical Scope	Last update	2018	2019	2019	2018
	Temp scope	N/A	N/A	N/A	N/A
	Geo scope	Madrid Region	City	Flanders	Thessaloniki Region
Gran. & Level of Detail	Temp gran	N/A	N/A	N/A	N/A
	Geo gran	Road/Street geometry	Road/Street geometry	Road segment	Road segment
	Level of Detail	Geometry, Names and Road Categories	Geometry, Names and Road Categories	Geometry, Name, Category, traffic characteristics	Geometry, Name, Category, detailed traffic characteristics
Availability & Accessibility		Whole project SHP All partners	Whole project SHP All Partners	Whole project SHP All partners	Whole project .net/SHP/KML/PNG/ 3DS/OBJ/OSG/IVI Spec. partners (NDA)
Relevance/Usability		Low	Low	Medium	High

Table 3. General overview of the Data Quality Assessment of Transport Network

4.2.3 Identified issues

The data sources available for Madrid and Regensburg present similar problems. The most relevant is that both data sources do not have a topological data structure, but simply provide the geometries of the transport network without defining the road segments and the allowed connections between them. In addition, the information they provide is very basic (e.g. name and road category). In the case of Regensburg, it should be added that the geographical scope is restricted to the city area.

As for Leuven, the data source Wegen Register is based on a topological data structure, and the only issue it presents is that the level of detail of the database is medium with little information regarding the traffic characteristics of the road segments.

The data source available for Thessaloniki does not present any relevant problems.

4.2.4 Relevance/Usability

The relevance of the available data sources for Madrid and Regensburg is low. As they do not have a graph data structure, their potential use within MOMENTUM is very limited to specific queries about the location and general characteristics of some locations in the transport network.

As for Wegen Register, the road database available for Leuven, its relevance is medium. The possible potential uses within MOMENTUM would be the modelling of the road network supply in Leuven, although the lack of relevant information on traffic characteristics of the road segments makes a great effort necessary to complete this information.

Finally, the relevance of the data source available for Thessaloniki is high because of its good geographical scope, timeliness and high level of detail. The potential use of this data source for MOMENTUM is the modelling of the road network and public transport network supply in this city.

4.3 Taxi Service Supply Data

4.3.1 Introduction

No relevant data sources within this category are available for **Madrid, Regensburg and Leuven**.

The only case study with relevant taxi supply (and demand) data is **Thessaloniki.** The dataset consists of records about taxi trips done in the city of Thessaloniki, Greece. This dataset's records are produced after processing the spatiotemporal taxi vehicle pulses recorded by onboard receivers of the Global Navigation Satellite System.

All information for each realized trip from 2016 onwards is recorded, including the coordinates and timestamp of the trip start and end, the temporal duration and travelled distance. The raw taxi data are anonymized and published by the Hellenic Institute of Transport, Center for Research and Technology Hellas (CERTH-HIT). The taxi fleet that generates the raw dataset operates under the name "Taxiway" taxi association, which group around 1000 taxis (around half of the taxis in Thessaloniki), and is the owner of the data. Taxiway has signed an MoU with CERTH-HIT, according to which CERTH-HIT has access to the data and is allowed to conduct processing and re-distribution, under certain conditions. The dataset will also be accessible to other specific partners under an NDA agreement. For each taxi trip recorded in the system, the data provided is the following: timestamps and GPS coordinates of the trip start and end locations, a flag indicating that the trip occurred with a passenger, trip duration in seconds and trip distance in meters, driver anonymized id, and vehicle anonymized id. These last two fields will only be available for CERTH because they have a high sensitivity to privacy. The dataset will be provided as a CSV file.

4.3.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		N/A	N/A	N/A	Taxi trips
Reliability		N/A	N/A	N/A	Medium / High
Temporal & Geographical Scope	Last update	N/A	N/A	N/A	Continuous update
	Temp scope	N/A	N/A	N/A	2016 - 2020
	Geo scope	N/A	N/A	N/A	City and suburbs/exurbs
Gran. & Level of Detail	Temp gran	N/A	N/A	N/A	Secondly
	Geo gran	N/A	N/A	N/A	Coordinates of trip origin/destination
	Level of Detail	N/A	N/A	N/A	Trip
Availability & Accessibility		N/A	N/A	N/A	Whole project CSV Spec. partners (NDA)
Relevance/Usability		N/A	N/A	N/A	High

Table 4. General overview of the Data Quality Assessment of Taxi Service Supply Data

4.3.3 Identified issues

One issue of this data source is that it only provides a partial view of the taxi supply in Thessaloniki. However, given the big size of the sample (around 50% of the taxis in Thessaloniki) and that no biases have been identified in the sample, it does not have a relevant impact.

4.3.4 Relevance/Usability

The relevance of the described data source for this project is high. It presents very good temporal and geographical scopes and granularities, and timeliness. Although it also contains taxi demand information, since it provides information at trip level (exact origin and destination coordinates and timestamps, as well as occupancy), it is possible to determine the spatial and temporal distribution of taxi supply in the Thessaloniki urban and metropolitan area. This data source can be used for the modelling of taxi supply in Thessaloniki, as well as a complementary data source for taxi demand analysis in combination with the one described in Section 5.10.

4.4 Car/Moto-Sharing Data Supply

4.4.1 Introduction

At the moment of writing this deliverable, no data source within this category is available for **Madrid** and **Leuven**. However, both EMT and Leuven's City Council have been and are in contact with different car/moto-sharing operators in their cities in order to have this type of data sources available for the two case studies.

In the **Regensburg Case Study**, the data comes from two public station-based electric car-sharing systems, one called das Stadtwerk.Earl that is deployed in the city of Regensburg and run by the company das Stadtwerk Regensburg Mobilität GmbH (a wholly-owned subsidiary of the city of Regensburg), and another one called KERL, deployed in the district of Regensburg and run by the company Kommunale Energie Regensburger Land eG. This is a small service with eight stations in the city of Regensburg and another ten in the district of Regensburg. The supply data for this data source is publicly available in Stadwerk.Earl webpage[1] but it has been also provided by SMO in CSV with information about the vehicle model and name, region, address, coordinates and start date of the car-sharing station.

Regarding **Thessaloniki Case Study**, the data source is obtained from free-floating e-scooter operators in Thessaloniki, concretely Hive and Lime. The data started to be collected by CERTH-HIT at the moment of writing this deliverable (February 2020). Therefore, it is already accessible by CERTH but it will be also accessible to other specific partners under an NDA. The dataset consists on the GPS coordinates of the locations of the currently available Lime and Hive scooters in the City of Thessaloniki. It is continuously updated at intervals of 5 to 10 minutes. The data will be provided in CSV format.

4.4.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		N/A	Station-based Car-sharing (locations and car info.)	N/A	Free-floating e-scooter
Reliability		N/A	High	N/A	High
Temporal & Geographical Scope	Last update	N/A	2019	N/A	Continuous updates
	Temp scope	N/A	2017-2019	N/A	Since February 2020
	Geo scope	N/A	City and district of Regensburg	N/A	City of Thessaloniki and suburbs/exurbs

[1] https://www.heyearl.de/e-carsharing/?L=0

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Gran. & Level of Detail	Temp gran	N/A	Monthly	N/A	5 to 10 minutes
	Geo gran	N/A	Station Level	N/A	Exact location
	Level of Detail	N/A	N/A	N/A	Individual e-scooter
Availability & Accessibility		N/A	Whole project Excel All partners	N/A	Whole project CSV Specific partners (NDA)
Relevance/Usability			High		High

Table 5. General overview of the Data Quality Assessment of Car/Moto Sharing Data Supply

4.4.3 Identified issues

No relevant issues were found for the two data sources available. The only issue regarding Thessaloniki's data source it is the fact that it has started to be collected very recently that will limit its temporal scope and therefore, the realization of historical analysis.

4.4.4 Relevance/Usability

The relevance of these car-sharing supply data sources for MOMENTUM is high. The potential uses within MOMENTUM are analogous to previous data sources. On the one hand, it can be used for the modelling of the car-sharing supply in **Regensburg** and free-floating e-scooter in **Thessaloniki**, and as a complementary data source for the ones described in Section 5.8 for analysing car-sharing and e-scooter demand in Regensburg and Thessaloniki, respectively.

4.5 Bike-sharing Data Supply

4.5.1 Introduction

In the **Madrid Case Study**, the bike-sharing supply data is provided by the EMT that is the company who also run BiciMAD, the public station-based bike-sharing service available in Madrid. The data are publicly available in its open data portal[1]. In this portal, for each month since July 2018, we have a zip file which in turn contains a JSON file with hourly snapshots on the situation of the bike-sharing stations. Concretely, this JSON file has a row for each snapshot containing the timestamp and the next information for each of the bike-sharing stations: a flag indicating whether the station is activated or not, name of the station, reservation counter, total number of bases, number of free bases, GPS coordinates and address of the station, number of available bases, number of docked bikes and stations id. The last month with data is June 2019. According to the EMT, the update frequency of this data is not fixed and depends on the availability of the computational resources for this process, as it is expensive.

[1] https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)

Regarding **Leuven Case Study**, the available dataset also comes from a station-based bike-sharing system run by the private company Blue Bike. It is a small-scale bike-sharing service with only one station in Leuven's train station. In this case, the data has a private character and will be only accessible for its use within the project. The available information, in Excel format, corresponds to the yearly number of available bikes from 2015 till 2019.

For the **Thessaloniki Case Study**, the bike-sharing supply data also comes from a public station-based service run by the private company Thessbike under a concession agreement. The provided data is private and at the moment of writing this deliverable, only accessible by CERTH, although it will be open to specific partners under an NDA agreement. It can be accessed by FTP/File download from a remote server or physical copy of a CSV file. The dataset contains a table with the name of each station, its coordinates in WGS84 and the station's capacity (maximum number of bicycles that can be stored). Furthermore, there is another file that provides temporal bike availability at each station (number of available bikes to be rented) every 10 minutes. Data was started to be collected at the beginning of 2019.

4.5.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Station-Based Bike-Sharing System (stations locations, bike and dock avail.)	N/A	Station-Based Bike-Sharing System (station location, bike avail.)	Station-Based Bike-Sharing System (stations locations, bike and dock avail.)
Reliability		High	N/A	High	High
Temporal & Geographical Scope	Last update	June 2019	N/A	December 2019	Continuous update
	Temp scope	Jul 2018 – Jun 2019	N/A	2015-2019	Since April 2019
	Geo scope	City centre	N/A	City	City
Gran. & Level of Detail	Temp gran	Hourly	N/A	Yearly	10 Minutes
	Geo gran	Station level	N/A	Station Level	Station Level
	Agg. Level	Aggregated	N/A	Aggregated	Aggregated
Availability & Accessibility		Whole project JSON Public	N/A	Whole project Excel All partners	Whole project Excel Specific partners
Relevance/Usability		High	N/A	Medium/Low	High

Table 6. General overview of the Data Quality Assessment of Bike Sharing Data Supply

4.5.3 Identified issues

Regarding **Madrid and Thessaloniki Case Study**, the main issue is the lack of information about other bike-sharing services in this city, especially those related to free-floating whose characteristics are somewhat different. However, at the moment of writing this deliverable, CERTH reported the availability of data from a free-floating bike-sharing system in Thessaloniki run by the private company Brainbox.

In the **Leuven Case Study**, the temporal granularity of the data is very coarse since we only have the number of available bikes each year.

4.5.4 Relevance/Usability

In the **Madrid and Thessaloniki Case Studies,** the relevance of the supply data available is high. The geographical and temporal scopes and granularities are good, although in the case of Thessaloniki is more limited. The potential uses of this data sources within this project are the modelling of the bike-sharing supply in Madrid and Thessaloniki, respectively, and the analysis of bike-sharing demand as a complementary data sources for those described in Section 5.3.

The relevance for the bike-sharing supply dataset available in **Leuven** is low. Its usability for this project is very restricted, and the only possible application is to complement the bike-sharing demand analyses in this city that may require this data (e.g. yearly variation of bike-sharing trips).

4.6 Parking Data Supply

4.6.1 Introduction

In the **Madrid Case Study,** the parking supply data is provided by the EMT, and it covers the fourteen public off-street parking run by this company. The data source consists of a public REST API[1] that gives supply information about these parking facilities in JSON format. For each parking, it provides the following information related to supply: parking category, name, schedule, type (e.g. rotation), information about the accesses available for the parking both by car and walking (GPS coordinates of the location, address, and a flag indicating whether it is a vehicle or pedestrian access), pricing in euros, address and a list with the facilities of the parking. The geographical scope is the city of Madrid.

The other case study with data about parking supply is **Leuven**. The data provider, in this case, is the City of Leuven, and it contains information about forty-three public off-street and shop&go parking spaces in Leuven. The dataset is private, but it will be accessible for all project partners during the whole project. It was provided in JSON format and contains the following information: type, location zone (e.g. city centre), pricing, capacity, accessibility facilities (e.g. elevator), schedule and GPS coordinates of the location. At the moment of writing this deliverable, Leuven also commented about the possibility of having supply information for on-street parking from May 2020. The data collection will be done for the whole territory of Leuven, including the outer city districts. They intend to map the parking spaces on 420 km of Leuven roads by using mobile mapping. Concretely, 360° images and point clouds will be collected by three vehicles and then used for making an inventory of on-street parking places. Each individual parking space, with correct dimensions, orientation and signalization will be drawn into the GIS.

No information within these categories is available for **Regensburg** and **Thessaloniki** case studies. However, at the moment of writing this deliverable, in Thessaloniki, we will explore the possibility of obtaining parking demand data through a data source that CERTH-HIT has been collected about parking status and driver's behaviour on major roads of the city of Thessaloniki.

[1] https://apidocs.emtmadrid.es/#api-Block_5_PARKINGS-parking_availability

4.6.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Public off-street parking lots	N/A	Public off-street and shop&go parking lots	N/A
Reliability		High	N/A	High	N/A
Temporal & Geographical Scope	Last update	December 2019	N/A	December 2019	N/A
	Temp scope	Real-time	N/A	2019	N/A
	Geo scope	City	N/A	City	N/A
Gran. & Level of Detail	Temp gran	Minutely	N/A	Yearly	N/A
	Geo gran	Parking facility	N/A	Parking facility	N/A
	Level of Detail	Aggregated occupancy	N/A	Aggregated occupancy	N/A
Availability & Accessibility		Whole project JSON All partners	N/A	Whole project JSON All partners	N/A
Relevance/ Usability		High		High	

Table 7. General overview of the Data Quality Assessment of Parking Data Supply

4.6.3 Identified issues

The main issue identified for the two data sources is the lack of supply data about private and semi-private off-street parking places, so the view of the parking supply in both cities is partial. Another limitation is the lack of historical information in both cases, which may affect some parking demand analysis.

4.6.4 Relevance/Usability

The relevance of both data sources for MOMENTUM is high. Although they only provide a partial view of parking supply in both cities, they do have very detail information for the supply of public off-street parking lots, especially in Madrid. Similarly to previous supply data sources, the potential uses within MOMENTUM are the modelling of parking supply in Madrid and Leuven, respectively, and the analysis of parking demand as a complementary data source to those that will be described in Section 5.12.

5 Data Quality Assessment for Transport Demand Data Sources

5.1 Public Transport Smart Card Data

5.1.1 Introduction

For the **case study in Madrid**, the data source is provided by the EMT, and it contains the **validations made by smart card owners** when they accessed the municipal buses. It is a private data source that in the context of the MOMENTUM project, it will be available only for some specific partners under an NDA agreement. The data will be provided in Excel or CSV format, and it will contain the following information: anonymised card id (one per user), bus validation time, line and bus stop, UTM coordinates, profile (child, youth, adult, elderly) and a value that indicates whether the validation was wrong or not (a value higher than 80 stands for a wrong validation). According to EMT, the data source is updated after every driver's shift, and they keep registers since 2006.

In the rest of the case studies, no such data is available. Regarding **Leuven**, the City Council studied the possibility of having data from the smart card MOBIB, which in this city is used by several public transport services, including De Lijn (bus) and NMBS (train). However, De Lijn, the company that Leuven contacted to obtain these data, reported that the data they had available was not reliable enough due to the incorrect use the passengers do of the smart card. As for **Regensburg and Thessaloniki**, the reason is that both cities have not a smart card for their public transport services.

5.1.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Smart Card Validations for urban buses	N/A	N/A	N/A
Reliability		High	N/A	N/A	N/A
Temporal & Geographical Scope	Last update	November 2019	N/A	N/A	N/A
	Temp scope	2006-2019	N/A	N/A	N/A
	Geo scope	City	N/A	N/A	N/A

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Gran. & Level of Detail	Temp gran	Validation time	N/A	N/A	N/A
	Geo gran	Bus stop	N/A	N/A	N/A
	Level of Detail	Anonymized Card ID	N/A	N/A	N/A
Availability & Accessibility		Whole project Excel, CSV Specific partners under NDA	N/A	N/A	N/A
Relevance/ Usability		High			

Table 8. General overview of the Data Quality Assessment of Public Transport Smart Card Data

5.1.3 Identified issues

The main issues identified for the data source available in the **Madrid Case Study** are the following:

- The data available is limited to urban EMT bus network. We will not have access to similar data from other public transport modes of the Madrid Region like commuter trains, metro or intercity buses.
- Smart card data users do not have to use their smart card when they get off the buses. This will require the use of different algorithms and models to solve this issue.
- Children younger than four years old do not need to pay in this service, so no data from them is collected.
- Users with a single ticket are not registered by this system which means that there are no records for occasional users.

5.1.4 Relevance/Usability

The relevance of the **smart card validations** for the case study **in Madrid** is high, because of the good geographical and temporal scope and granularity, and mainly because it provides longitudinal traceable information about the users since the anonymised card ID identifies a specific smart card over time. The potential uses of this data source within the MOMENTUM project are the analysis of travel behaviour and activity patterns of EMT buses' users and the study of the demand of this service (e.g. variation over time, trip length distribution, etc.); the estimation the modal split in the Madrid city area for EMT buses and the travel times using this service. Another interesting potential use for MOMENTUM is the study of the influence of PT regularity on PT demand.

5.2 Other Public Transport Demand Data

5.2.1 Introduction

For the **Regensburg Case Study**, there is one data source regarding PT demand that it is not related to smart card. This data is collected and owned by the Stadtwerk Regensburg Mobilität GmbH (SMO), and it contains passenger counts data that are extracted from sensors installed in the buses. SMO reported that around 10% of the buses in Regensburg are equipped with these sensors. This data source will be available during the whole project, and it is

extracted from FAN Software as Excel or CSV files. For each line with an equipped bus, the data source provides the following information: whether the aggregation period is a specific day and/or hour or, alternatively, a weekday or group of weekdays; average daily number of passengers in that line; reference period for which the data are aggregated; file creation date; number of collected trips during the aggregation period (and therefore, used to obtain the aggregation measurements); and finally, (the main part of the data source) for each bus stop it provides the following information for the two-line directions (the data is being collected since 2017, and the last update was on November 2019):

- Bus stop name
- People entering the bus and people leaving the bus
- Total number of people entering/leaving
- Number of passengers already in the bus
- Sum of people already on the bus plus people entering the bus
- The percentage of people that took the bus at that stop w.r.t the total number of passengers that took the bus in all bus stops of that line direction.

No data sources are available within this category in **Madrid, Leuven** and **Thessaloniki** case studies.

5.2.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		N/A	Passenger counts from sensorized buses		
Reliability		N/A	Medium / High	N/A	N/A
Timeliness & Geo Scope	Last update	N/A	November 2019	N/A	N/A
	Temp scope	N/A	2017 – 2019	N/A	N/A
	Geo scope	N/A	City	N/A	N/A
Gran. & Level of Detail	Temp gran	N/A	Hourly	N/A	N/A
	Geo gran	N/A	Bus stop	N/A	N/A
	Level of Detail	N/A	Aggregated counts	N/A	N/A
Availability & Accessibility		N/A	Whole project Excel, CSV All partners	N/A	N/A
Relevance/Usability			Medium		

Table 9. General overview of the Data Quality Assessment of Other Public Transport Demand Data

5.2.3 Identified issues

The main issues that present the data source described above for the **Regensburg Case Study** are the following. First, that in some cases the accuracy of the data may be medium and that it provides aggregated data, that is, there is no traceable information about users nor information disaggregated by bus arrival event at a specific stop. Second, the software containing these data only allows for manual extraction of aggregated data for one period of analysis at a time. If a time granularity of one hour is required, a manual extraction should be made for each hour of the period to be studied (e.g. a week of data at hourly granularity would require 7x24 = 168 manual extractions). This makes the obtention of a dataset with a good temporal scope a laborious task.

5.2.4 Relevance/Usability

Bus passenger counts data for **Regensburg** has medium relevance for MOMENTUM because of the limitations presented above. Taking into account that it is hourly aggregated information and that there is no possibility of tracking users, the potential uses of this data source for MOMENTUM are the analysis of the aggregated demand of bus transport in Regensburg in different weekdays or periods (e.g. variation over time, the spatial distribution of the demand, etc.). Another interesting possibility is the sensitivity of this demand to exogenous factors as the weather. It should finally be noted that only 10% of buses are equipped with these sensors and that the reliability of these sensors is medium/high which means that the conclusions of these analyses are limited to the lines and routes taken by these buses and should account for the possible measurement error. The joint study of this data source with the one for PT time deviations in Regensburg (see Section 4.1) can also be useful to study the influence of PT regularity on demand for this mode of transport.

5.3 Bike-sharing Demand Data

5.3.1 Introduction

In the **Madrid Case Study**, the bike-sharing demand data is provided by the EMT that is the company who also run BiciMAD, the public station-based bike-sharing service available in Madrid. The data are publicly available in its open data portal[1]. In this portal, for each month since April 2017, we have a zip file which in turn contains a JSON file with data on the use of this service. Concretely, this JSON file contains the following information: trip id, anonymised user code per day, unplug station and base ids, plug station and base ids, unplug hour time, travel time and track (a GeoJSON with the GPS trajectory followed by the user in this trip). The last month with available data is June 2019. According to the EMT, the update frequency of this data is not fixed and depends on the availability of the computational resources for this process, as it is long and expensive.

Regarding **Leuven Case Study**, the available dataset also comes from a station-based bike-sharing system run by the private company Blue Bike. It is a small-scale bike-sharing service with only one station in Leuven's train station. In this case, the data has a private character and will be only available for its use within the project. The available data corresponds to the statistics of use the company sends to the City Council every year. The data is provided in Excel format, and it contains the following information: demand (number of rides) per year/month/week, total number of bike-sharing users with at least one ride, mean number of rides per user, number of bike-sharing users living in Leuven, mean number of total rides (anywhere) per user, origin of users, postcode of active users in Leuven and some global statistics about the impact of this service on mobility in the city of Leuven. There is historical data about the number of trips from 2015 but with a monthly temporal granularity and the weekly number of trips but only for 2018.

[1] https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)

For the **Thessaloniki Case Study**, the bike-sharing demand data comes from a public station-based service run by the private company Thessbike under a concession agreement. The data source is private, and at the moment of writing this deliverable, it is only accessible by CERTH, but it will be open to some specific partners under an NDA agreement. It is provided in CSV format, and it contains the following information per bike-sharing trip: unique rental identifier, unique user identifier, unique bike identifier, timestamp of rental start, timestamp of rental end, start station id, end station id and total rental cost. It has been collected since 2014, and it is minutely updated.

In the case of **Regensburg,** there is no data source within this category because the city does not have a bike-sharing service yet.

5.3.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Station-base bike-sharing	N/A	Station-base bike-sharing	Station-base bike-sharing
Reliability		High	N/A	High	High
Temporal & Geographical Scope	Last update	June 2019	N/A	2018	Continuous update
	Temp scope	2017 - 2019	N/A	2015 - 2018	Since 2014
	Geo scope	City Center	N/A	Only one bike haring point in the train station	City
Gran. & Level of Detail	Temp gran	Hourly	N/A	Weekly	Minutely
	Geo gran	Individual level	N/A	Bike-sharing station	Exact station of the start and completion of a bike rental
	Level of Detail	Trip with anonymized id per user and day	N/A	Aggregated number of trips	Trip with unique id per user
Availability & Accessibility		Whole project JSON Public	N/A	Whole project excel All partners	Whole project CSV Specific partners subject to NDA
Relevance/Usability		High		Medium	High

Table 10. General overview of the Data Quality Assessment of Bike-Sharing Data Demand

5.3.3 Identified issues

The main issue identified in the demand data from the BiciMAD service in **Madrid** is the anonymisation performed to the trip information. Since we have one anonymised user ID per day, it makes it not possible to trace users longitudinally. To a lower extent, another issue is the truncation of the bike unplugged timestamp to the closest lower o'clock hour (in order to ensure the privacy of the data) which may reduce the accuracy of the analyses to an hourly granularity. However, it is not expected to be relevant in the context of MOMENTUM.

The data from the Blue Bike service in the **Leuven** is the one that presents the most relevant limitations. The most important ones are the low temporal granularity of the data (weekly), the small deployment of this service in the city with only one bike-sharing point at the train station and the aggregated character of the data with no possible tracking of users. The main reason behind the first issue mentioned is the reluctance of the company to share more detailed data of its service with the City Council.

In the case of **Thessaloniki**, the only issues found, although not relevant, are the reduced deployment of this service in Thessaloniki with only eight docking stations and the lack of socio-demographic information about the users such as age range, gender or home location. However, regarding this last issue, at the moment of writing this deliverable, CERTH reported that there are a mobile application launch and a large promotion campaign scheduled for March and April 2020. On it, they will gather data about the system users as age, sex, home location (probably as zip code), etc.

5.3.4 Relevance/Usability

The relevance for MOMENTUM of the bike-sharing data available for **Madrid and Thessaloniki Case Studies** is high. In both cases, information at trip level with user traceable information allows a disaggregated analysis of mobility patterns of people that utilise this transport mode. The potential applications of these data sources within MOMENTUM are the analysis of the demand and use patterns of bike-sharing, including aspects such as the obtention of OD Matrices, distribution of trip distances, travel times, the variation of the number of users/trips over time, etc. In the case of Madrid, since the longitudinal traceability of the users across different days is not possible, it limits some studies such as the inference of the activity patterns associated with trips. However, this data source will allow the analysis of the factors influencing the adoption of bike-sharing in this city by the combination of the described demand data with the socio-demographic data sources available (see Section 7) through the postal code of the user's residence. In the case of Thessaloniki, this type of analyses will be more limited since no socio-demographic information about the user is available.

In the **Leuven Case Study**, the relevance of the data source available is medium, since it only allows analyses at an aggregated level and restricted to the number of trips given that no other information is available. In this case, the potential uses of this data source within MOMENTUM are the analysis of the weekly trends in the use of this bike-sharing service and its dependence from exogenous factors as weather or calendar. These analyses should bear in mind that use of this service may be biased towards destinations near this location and probably towards users who make frequent use of the train. However, this last aspect makes it interesting for MOMENTUM, since it may allow studying the complementarity of this shared-transport mode with another means of PT such as the train. To cite an example, according to data provided by the company to the City Council, 22% of trips on PT + share-bike come from trips that were previously done by car.

5.4 Cycling Data

5.4.1 Introduction

In the **Madrid Case Study**, the cycling demand data come from the City Council, and it is publicly available in its open data portal[1]. It is collected by 33 fibre-optic sensors located in 19 different locations of the city. The data are provided in Excel format and contain information relative to the number of bicycles counted by each sensor at intervals of 15 minutes. It started to be collected on January 1st, 2019, and it is updated every three months. However, at the time of writing this deliverable, the open data portal reported an incidence in the third trimester of 2019. The data relative to July, August and September 2019 will not be published because of a technical failure in the data collection system which led to a significant decrease in the counts of all permanent sensors after a recalibration of the measuring points.

For the **Leuven Case Study**, the cycling demand data also belongs to City Council, but, in this case, its access is restricted to MOMENTUM partners, and it can only be used for project purposes. In this case, there are only five sensors in five different locations in the city area. The data, available in Excel format, contains the number of bicycles detected at the sensor location at each hour and at each cycling direction as well as the sum of the counts in all directions. The temporal scope of the available data ranges from 2015 till 2017. There is no more recent data because these sensors are broken since 2017. Leuven reported that these sensors will be working again in 2020 and that 15 more sensors will be installed but is it not clear whether this data will be available for its analysis in the timeframe of WP3.

In the case of **Regensburg**, this data is also provided by the City Council, but they are not publicly available. The access was granted only for project purposes as in the previous case. Unlike Madrid and Leuven, the cycling demand data available come from a frequency survey undertaken in 14 locations in the old town between September 17th, 2018 and September 22nd, 2018. This data source consists of manual counts taken from 8:00 AM till 8:00 PM in every day of the survey, aggregated hourly. Data was provided in Excel file format. In this case study, Regensburg also provided some cycling counts on different locations in the city that were carried out at specific days in several years. However, they were only available in PDF format and with a complex structure, so we decided to discard them because of the complexity for extracting the useful information.

In the case of **Thessaloniki,** there is no data source within this subcategory.

5.4.2 General overview of the Data Quality Assessment

Type & Data quality dim.	Madrid	Regensburg	Leuven		Thessaloniki
Type	Counts from fibre-optic sensors	Manual Counts	Counts from fibre-optics sensors	Counts from Crowdsourced Camera sensors	N/A
Reliability	Medium	Medium	High	Medium	N/A

[1] https://datos.madrid.es/portal/site/egob

Type & Data quality dim.		Madrid	Regensburg	Leuven		Thessaloniki
Temporal & Geographical Scope	Last update	September 2019	2018	2017	2019	N/A
	Temp scope	Jan-Jun 2019	6 days in September 2018	2015-2017	2019	N/A
	Geo scope	City (19 locations)	Oldtown	City (5 locations)	City	N/A
Gran. & Level of Detail	Temp gran	15 min	Hourly	Hourly	Hourly (only on daytime)	N/A
	Geo gran	Sensor locations	Counting locations	Sensor locations	Sensor locations	N/A
	Level of Detail	Aggregated counts	Aggregated counts	Aggregated counts	Aggregated counts	N/A
Availability & Accessibility		Whole project CSV/Excel Public	Whole project PDF/Excel All partners	Whole project CSV All partners	Whole project JSON/CSV All partners	N/A
Relevance/Usability		Medium	Low	Medium	Medium	N/A

Table 11. General overview of the Data Quality Assessment of Cycling Data

5.4.3 Identified issues

In the **Madrid case study,** the main issue is the low number of sensors available considering the size of the city. Furthermore, because of the incidence reported above, there will not be data for July, August and September, that is, there is no data available for the summer period.

Regarding the **cycling counts by fibre-optic sensors** in **Leuven**, the major issue found is the shortage of sensors, as there are only five measuring points spread over the city. The lack of data from the last two years may lead to skipping some recent trends in cycling demand in Leuven, but it is not a relevant issue.

As for the **Telraam** data source for bikes, it also presents relevant issues, most of them related to the device used for bicycle detection. Although the number of cameras is high, its distributions not geographically homogeneous with some areas with a high density of detection devices whereas others without sufficient coverage. Besides, the devices only work during the daytime because the technology is based on a camera and artificial vision methods, that does not work well in poor light conditions. Furthermore, the device becomes inactive 20-35% of the time, while automatic calculations and calibrations are performed. Although this downtime can be corrected using aggregated data and imputation methods, it decreases the accuracy of this detection system for bikes. Finally, depending on the inclination and orientation of the cameras towards the street, the device may only be able to detect bicycles passing through one of the two sidewalks so that it may skip an important number of bikes.

Finally, the cycling demand dataset available in **Regensburg** is the one that presents more problems. First, the bicycle counts are relative to a very short time (only six days). Second, the measurements were manually taken which is usually prone to errors, especially if we consider that these counts have been done for long periods (12 hours a day). And third, there were only 14 locations for the manual counts of bikes.

5.4.4 Relevance/Usability

Regarding **Madrid's Case Study**, the relevance of the cycling demand data source is medium. The low number of measuring locations and their irregular coverage of the city make it potentially usable only for capturing general trends in the cycling demand of Madrid (e.g. temporal variability) or for local analyses at the specific location of the sensors. Another potential use, very interesting for MOMENTUM, is the joint analysis of this data source with the bike-sharing data from BiciMAD (see Section 5.3) and the weather data for Madrid (see 6.2) to study how weather conditions may influence cycling demand in this city. The short temporal scope of this data source (January-June 2019) limits the historical analyses and the variability in weather conditions (e.g. summer). However, this issue may be alleviated as more data become available during 2020.

In the case of cycling counts from fibre-optic sensors and Telraam in **Leuven**, the relevance and potential uses in MOMENTUM are analogous to the previous data source. The only particularities to highlight are the lower number of sensors (five) but a better temporal scope of the first dataset mentioned. This will improve its usability for historical analyses and the study of the influence of weather conditions.

Finally, in the **Regensburg** case study, the relevance of the available data source is low. Therefore, the usability of this data source within MOMENTUM is very limited. It may help to know the variations of the cycling demand in Regensburg's oldtown along different hours or weekdays or to identify peaks times, but the conclusions must be taken with caution because the risk of bias is very high given the very short temporal scope.

5.5 Pedestrian Data

5.5.1 Introduction

In the **Madrid Case Study**, the data source available for pedestrians is analogous to that described for bicycles in the previous section, except that in this case, there are 19 sensors in 19 different locations. The data are also open and available on the same portal.

Regarding **Regensburg**, the data source is the same as depicted in the previous section, since pedestrians and cyclist were counted simultaneously during the frequency survey. Similarly, to the last data source, Regensburg also provided pedestrian counts on different locations in the city that were carried out at specific days on several years. We decided to discard them for the same reasons explained above.

In the case of **Leuven**, there are again two data sources. The first one is Telraam, given that the described devices detect pedestrians too. The second one corresponds to hourly pedestrian counts coming from WiFi detectors deployed in 23 different locations of the city. The data is private and belongs to the City Council. So far, there is only one month of data (November 2019), and the data source contains for every hour of this period the pedestrian counts in the locations above.

No data source is available within this category for **Thessaloniki**.

5.5.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven		Thessalon
Type		Counts from Fiber-optic sensors	Manual Counts	Counts from Crowdsourced Camera sensors	Wifi detectors	N/A
Reliability		Medium	Medium	Medium	Medium	N/A
Temporal & Geographical Scope	Last update	June 2019	2018	December - 2019	Continuous update	N/A
	Temp scope	2019	6 days in September 2018	2019	Since 2017	N/A
	Geo scope	City (19 locations)	Old town	City	City Center	N/A
Gran. & Level of Detail	Temp gran	15 mins	Hourly	Hourly	Hourly	N/A
	Geo gran	Sensor location	Count location	Sensor location	Sensor location	N/A
	Level of Detail	Aggregated counts	Aggregated counts	Aggregated counts	Aggregated counts	N/A
Availability & Accessibility		Whole project csv / excel All partners	Whole project csv / excel All partners	Whole project JSON All partners	Whole project JSON All partners	N/A
Relevance/ Usability		High	Low	High	High	

Table 12. General overview of the Data Quality Assessment of Pedestrian Data

5.5.3 Identified issues

In the cases of **Madrid, Regensburg and Telraam data for Leuven**, the issues found would be the same as those mentioned in Section 5.4.3.

As for the **pedestrian counts from Wifi sensors in Leuven**, the main limitations come from the short temporal scope (only one month) and for the low number of sensorized locations. Regarding the first issue, Leuven reported the possibility of having more data in the near future. To a lower extent, another limitation of this data source is the reliability of these sensors, which is usually medium/high.

5.5.4 Relevance/Usability

The three data sources shared with cycling demand data have analogous usability to the one explained in Section 5.4.4.

Regarding the **pedestrian counts based on Wifi sensors from Leuven**, as long as no more data is available, the relevance for MOMENTUM is low, and the potential uses would be similar to that of the data source of Regensburg.

5.6 Mobility Surveys

5.6.1 Introduction

Starting with the case study in **Madrid**, the available data source is the 2018 Household Survey for the region of Madrid, but it will not be available until March 2019. The results of the survey, which was presented at the moment of writing this deliverable (February 26th, 2019), are expected to be accessible as open data soon.

In the **Regensburg Case Study,** the mobility survey available corresponds to the time series study "Mobility in cities – SrV" [1] performed by TU Dresden in 2018, whose results were available for Regensburg at the end of 2019. This is the eleventh edition in a series of studies that began in 1972 that uses a household survey to obtain data on traffic behaviour in selected cities and regions of the Federal Republic of Germany with Regensburg among them. For Regensburg, this survey took place from February 2018 till January 2019 on middle working days (Tuesday, Wednesday and Thursday), and only households in the Regensburg urban area were surveyed. A total of 2501 people in 1116 households were successfully interviewed. For this study, the Regensburg urban area was divided into five sub-areas: centre, north, south, east and west. The survey contained questions related to emerging mobility solutions, concretely, about the general use of car-sharing or bike-sharing. Regensburg obtained from TU Dresden access to the micro-data of the survey for Regensburg region and they will be available for all project partners. The data is stored in Microsoft Access format.

For the **Leuven Case Study**, four relevant surveys are available on mobility issues. The first one, Stadsmonitor[2] (City Monitor), is specific for Leuven but it only considers people registered in this city. This survey is carried out every three years in different cities of the Flanders region in Belgium, including Leuven. It covers residents over 16 years old in order to collect different indicators on their perception of various issues related to the liveability of the city among which we can find mobility. Approximately 2400 residents of Leuven, divided in 8 different districts, were interviewed between May 2nd and June 6th 2017 on aspects such as place of residence, sex, number of members in the household, work status, income, nationality, amount and type of vehicles owned per household, PT passes in the household, frequency and modal choice for leisure trips, frequency and modal choice for travel to the workplace, etc. In addition, it contains questions regarding emerging modes of transport such as

[1] https://tu-dresden.de/bu/verkehr/ivs/srv/srv-2018
[2] https://www.gemeente-en-stadsmonitor.vlaanderen.be/

the number of subscriptions to car-sharing services at the household level, or the wish to use car-sharing services now or in the future. For this survey, Leuven has provided access to the micro-data in Excel format.

The second survey concerning mobility in Leuven is called OVG and is conducted by the Department of Mobility and Public Works of the Government of the Flanders Region. It is accomplished annually, and in the last edition (2018), 1595 people were successfully interviewed between mid-January 2017 and mid-January 2018. This survey considered a family questionnaire where it asked aspects such as the number and type of vehicles in the household (both owned and through other mechanisms as car-sharing or company car), income and number of members in the household; a personal questionnaire on aspects such as age, gender, living situation, income, driver license, studies, marital status, householder, frequency of use of different transport modes, work status, workplace location, frequency and modal choice for travel to the workplace; and a travel diary in which the interviewed listed the details of the different trips they made on the specific day indicated by the interviewer. Although the survey does not have specific questions about the use of new forms of mobility, it does include car-sharing as a possibility of having a vehicle at home and questions about the use of car-pooling to get to work. In this case, we only have access to the PDF file of the survey report with processed results that is publicly available.

The third survey is done explicitly for students by the KU Leuven. This survey consists of a questionnaire that is sent to all students. The last survey refers to 2017, were approximately 10% of students responded. It contained questions about level of studies; disabilities; bike ownership and reasons in case of no ownership; driver's license; ownership of a pass for students as well as the frequency and type of activity for which it is used; location, frequency and modal choice for trips to parent's home; frequency and modal choice for trips to classes; modal choice and reasons for trips to classes and other activities; use of bike rental (Velo); rank of most used bus lines in Leuven; reasons for which the student would use bus, car, bike or walk more often in Leuven. No specific questions about the use of emerging mobility solutions have been identified. Leuven has provided access to this survey's microdata to all partners, and it is available in CSV format.

The fourth survey in Leuven is called Diagnostiek woon-werkverkeer [1](Diagnosis of commuter traffic). The survey is organized every three years by the Federal Public Service Mobility and Transport in which participating companies and government institutions in Belgium that employ on average more than 100 employees. They are asked about the commuting trip of their employees. The most recent data collection was held between 1 July 2017 and 31 January 2018. During this period 3,951 different employers took part (accounting for 11,536 workplaces that employ more than 1.5 million people). The questionnaires completed by the employers contains questions about total number of employees; number of employees included in the survey; organization of working time; main transport modes for commuters; main transport modes for commuting according to postcode of employees' residence; possible modes of transport in the pre- and post-route; accessibility of the company premises and mobility problems related to parking lots, bicycle, public transport and car; mobility policy of the employer regarding the use of bicycle, collective transport, carpooling, car and telework. Leuven also provided access to the microdata of this survey to all partners only for project purposes. This data will be available in Excel format for the whole project.

Regarding **Thessaloniki Case Study**, the available travel survey took place during the period 2017-2018. The mobility survey was conducted to a sample of approximately 9.000 households from all Municipalities of the Prefecture of Thessaloniki. The ultimate number of questionnaires completed reached the number of 11.000, including about 2.000 questionnaires that addressed mobility-impaired persons. In parallel, a stated preference survey was conducted in 3.000 of the 10.000 households, using game-cards: alternative trip scenarios were developed, each one of them containing a different combination of trip characteristics (trip time, time to reach

[1] https://mobilit.belgium.be/nl/mobiliteit/woon_werkverkeer

the mode, time to reach the destination once leaving the mode, waiting time, trip cost, trip comfort). Three general sets of game-cards were developed, based on the Municipality of residence of the respondent.

The results of satisfaction surveys are also available for public transport users (approx. 2.000 questionnaires completed), bicycle users (approx. 500 questionnaires completed) and pedestrians (approx. 2.700 questionnaires completed). The survey included questions about the number of household members, number of household members older than 16, type of vehicle ownership, fuel used by the vehicle, annual household income, gender, employment status, hours of work, driving licence holder, access to private car, personal vehicle, type of private vehicle, number of daily trips, public transport card for unlimited use, level of education, annual personal income, trip characteristics (start trip numbering, start location, time of arrival to destination, type of trip activity, time of departure from destination, mode of transport used); The microdata of this survey is provided in Excel format. CERTH already have access to this data, and some specific partner will also have access to them under an NDA.

5.6.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Mad.	Regensb.	Leuven				Thessal.
Type		N/A	Household survey	City liveability survey	Student survey	Household survey	Work-place Survey	Household survey
Reliability		N/A	High	High	High	High	High	High
Temporal & Geographical Scope	Last update	N/A	2018	December 2019	2017	2017	2018	2018
	Temp scope	N/A	Feb 2018 – Jan 2019	May 2 – June 6 2017	2017	2017	July 1 2017 - Jan 31 2018	2017 - 2018
	Geo scope	N/A	City	City	City	Flanders	City	Prefecture of Thessalon.
Gran. & Level of Detail	Temp gran	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	Geo gran	N/A	Five sub areas: north, south, east, west and centre	8 sub-areas	Postal code	Three sub-areas: Brussels, Flanders and Wallonia	Company premises	Traffic zone
	Level of Detail	N/A	Micro-data	Micro-data	Micro-data	Aggregated results	Micro-data	Micro-data

Type & Data quality dim.	Mad.	Regensb.	Leuven				Thessal.
Availability & Accessibility	N/A	Whole project PDF All partners	Whole project CSV All partners	Whole project CSV All partners	Whole project PDF All partners	Whole project PDF All partners	Whole project Excel Specific partners
Relevance/ Usability	N/A	Medium	High	High	Low	High	High

Table 13. General overview of the Data Quality Assessment of Mobility Surveys

5.6.3 Identified issues

Given that most of the travel surveys available are cross-sectional household surveys, they share common issues to these data collection methods as sampling errors, measurement errors, non-response bias, processing errors, hypothetical bias, etc.

Particularizing for each case study, in **Regensburg,** the main issue is the low spatial granularity of the household survey as it only considers five sub-areas within the city with entails a high aggregation level.

As for the **Leuven Case Study**, each of the surveys presents its own shortcomings. Starting with City Monitor, the main issues found are three: 1) as the objective of the survey is more general than mobility, the information about this topic is limited; 2) the sub-areas in which the city is divided are coarse-grained which also implies a high level of aggregation; and 3) it does not contain information about students as, by Belgian law, they are registered at their parent's house. The OVG Flanders survey is a complete survey, but it has the limitations of providing aggregated data at a regional level and in addition, only the PDF is accessible, which, as mentioned above, makes it difficult to process it automatically. The only issue found in the Student Survey is the lack of questions about emerging mobility solutions. Finally, the WWV Survey is complete, but it also presents problems: it is only limited to commuters, and the results may be biased to the characteristics of those employees who answer the questionnaires.

Regarding the travel survey available for **Thessaloniki,** the main issue is no consideration in the questionnaire design of new emerging mobility solutions.

5.6.4 Relevance/Usability

For the sake of simplicity and given that, as we have already commented above, most travel surveys are household surveys, we list here the potential uses that this type of data source can have within MOMENTUM: the creation of travel generation and attraction models, analysis of the modal split and usage patterns of different transport modes, distribution of travel times and distances, study of the socio-demographic factors that influence the adoption of a particular mode of transport, etc.

Entering each particular case, in the **Regensburg Case Study**, the relevance of the household survey available is medium as the high level of aggregation of the data reduces its potential uses in MOMENTUM. Although one of the interesting aspects of this survey is that it contains questions on the use of car-sharing in Regensburg, as the use of this service is, for the moment, somewhat residual in this city, the number of respondents who provided information on this aspect was very low, making difficult to draw meaningful conclusions.

As for the case study in **Leuven**, all surveys have a high relevance except OVG, which has a medium/low relevance as it contains mainly regionally aggregated data. Starting with the City Monitor survey, its most interesting use within MOMENTUM is the analysis of factors influencing the ownership of a car-sharing service subscription or the willingness to use it now or in the future, as it contains questions in this respect. The potential applications of the Students Survey would be, on the one hand, to complement the previous survey, since it does not collect data on students in Leuven (they represent a third of the population of this city). On the other hand, since it is a complete questionnaire, to analyse the travel patterns and behaviour of the younger strata of the population (although it would be biased towards students) with respect to the classical modes of transport as walking, cycling, car or public transport. Finally, the WWV survey allows us to particularize the analyses to the commuters, and what is more interesting, to study the use and adoption they make of the different modes of transport, since we have aggregated data at the level of the commuter's residence and workplace postal codes, and information on companies' policies to encourage a more sustainable mobility of their employees.

5.7 Telecom Data

5.7.1 Introduction

This data source is only available for the **Madrid Case Study** and it comes from a private agreement between Nommon and Orange Spain. These data consist of call detail records (CDRs) and socio-demographic information about Orange clients and the geographic location of the telecommunication towers. The sample size of this data source is larger than the 20% of the Spanish population (in the day taken as reference for this deliverable, October 15th 2019, 18 million users were connected to the network). The temporal granularity of this data source depends on the user usage of the mobile phone, but it usually provides one register every 30 minutes. The analyses done with this type of data sources have generally up to an hourly temporal granularity. The geographical granularity of the data depends on the spacing between towers providing coverage in the area. In the city of Madrid, an analysis of the radius of coverage showed that in nearly 50% of the towers this radius is lower than 300 meters. Nommon has records of this data source from 2016 and geographical coverage considered for MOMENTUM will be the City of Madrid.

5.7.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Call Detail Records			
Acc & Prec		High	N/A	N/A	N/A
Temporal & Geographical Scope	Last update	2019	N/A	N/A	N/A
	Temp scope	2016-2019	N/A	N/A	N/A
	Geo scope	City	N/A	N/A	N/A

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Gran. & Level of Detail	Temp gran	20-30 mins	N/A	N/A	N/A
	Geo gran	Mobile network antenna	N/A	N/A	N/A
	Level of Detail	Individual users	N/A	N/A	N/A
Availability & Accessibility		Whole project CSV Only available for Nommon	N/A	N/A	N/A
Relevance/ Usability		High	N/A	N/A	N/A

Table 14. General overview of the Data Quality Assessment of Telecom Data

5.7.3 Identified issues

The main issues identified for this telecom data source available for the **case study in Madrid** the following:

- The geographical granularity of mobile phone data depends on the mobile network antenna density, giving spatial uncertainty of around 200m in big cities, but up to 2km in rural areas.
- User profile information may not always be accurate, as the client (who is the person that appears in the Mobile Network Operator files) may not be the user of the mobile phone (e.g. teenagers who have their phone paid by their parents, for example).
- When no network coverage maps are available, Voronoi areas are used as a proxy of antenna coverage, which it has potential to be improved (they do not take into account obstacles or antenna technology)

5.7.4 Relevance/Usability

The relevance of this data source for MOMETUM is high. Mobile phone data is a powerful data source to obtain information about mobility patterns of the population since the sample size is much bigger than compared to other data sources. Furthermore, both spatial and temporal resolution is good enough to characterise the mobility of users. The potential uses of this data source within MOMENTUM are the inference of common locations/activities of transport users (e.g. work, home, etc.); the obtention of general mobility models, population flows and travel patterns for people living in Madrid; the extraction of OD matrices for Madrid Region and the estimation of transport mode and route choices used by users. Furthermore, given the good temporal scope and relevance of this data source, another potential application is the analysis of the temporal and seasonal variability of mobility.

5.8 Car/Moto-Sharing Data Demand

5.8.1 Introduction

In the **Regensburg Case Study,** the data, that was collected by SMO, came from a public car-sharing system deployed in Regensburg City and Metropolitan Area that it is run by the semi-public company REWAG

Regensburger Energie- und Wasserversorgung AG & Co KG. The data are private, and they are only available for MOMENTUM project purposes. The data were extracted from the operating tool used for providing the customer service and for tracking the vehicles. It also contains cancelled bookings. The temporal scope of the dataset provided ranges from November 2016 to November 2019 and it includes the following information for each trip: start date, start time, return date, return time, used time, pick up location, return location, make, model, registration, unit, trip length, estimated trip length, start booking (km), return (km), type, status, booking edited, booked start date, booked start time, booked return date, booked return time and booked trip.

In the case of **Thessaloniki**, as explained in Section 4.4.1, the data source is obtained from free-floating e-scooter operators in Thessaloniki, concretely Hive and Lime. As supply data, it started to be collected by CERTH-HIT at the moment of writing this deliverable (February 2020). It is already accessible by CERTH, but it will also be available to other specific partners under an NDA. The dataset consists of the trips done with e-scooters from these providers in Thessaloniki. This dataset is extracted by processing the locations of available scooters and monitoring the change of location for each e-scooter. Concretely, by accessing an API from Lime and Hive, CERTH records the exact location and battery levels at specific timestamps. Since a unique static identifier is provided for each e-scooter, it is possible to process and extract the completed trips. The dataset will be accessible in CSV format.

No data source information is available for **Madrid and Leuven** case studies. However, as mentioned above, both EMT and Leuven's City Council have been and are in contact with different car/moto-sharing operators in their cities in order to have this type of data source available for their respective case studies.

5.8.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type			Station-based car-sharing		Free-floating e-scooter
Reliability		N/A	High	N/A	High
Temporal & Geographical Scope	Last update	N/A	2019	N/A	Continuous updates
	Temp scope	N/A	Nov. 2016 – Nov 2019	N/A	Since February 2020
	Geo scope	N/A	City and District of Regensburg	N/A	City of Thessaloniki and suburbs/exurbs
Gran. & Level of Detail	Temp gran	N/A	Trip/Booking Event	N/A	5 to 10 minutes
	Geo gran	N/A	Car-sharing station	N/A	Exact location
	Level of Detail	N/A	Trip (no information about users)	N/A	Trip at individual e-scooter (no information about users)

Type & Data quality dim.	Madrid	Regensburg	Leuven	Thessaloniki
Availability & Accessibility	N/A	Whole project CSV All partners	N/A	Whole project CSV Specific partners (NDA)
Relevance/Usability		High		High

Table 15. General overview of the Data Quality Assessment of Car/Moto-Sharing Data Demand

5.8.3 Identified issues

The main limitations of the car-sharing demand data available for **Regensburg Case Study** is, first, the complete absence of user information, both for traceability (user ID) and for socio-demographic characterization (e.g. age, gender, etc.) because of privacy issues; and second, the small size of the deployment of this service, with only seven stations in the city of Regensburg and 17 in its district, so its use is still residual compared to other transport solutions. Furthermore, SMO reported that could be inaccuracies in start time and return time, registration times, number of kilometres (trip) and type and status of the booking which may affect the trustability of the data.

In the case of **Thessaloniki**'s data source, the main limitation is shared with Regensburg, as no information about users is available. Another issue is the fact that it has started to be collected very recently that will limit its temporal scope and therefore, the realization of historical analysis. Due to the recency of this data source, an in-depth analysis of more issues could not be accomplished at the moment of writing this deliverable.

5.8.4 Relevance/Usability

Despite the previously mentioned issues, the relevance of these data sources for MOMENTUM is high, although, in the case of Thessaloniki, it should be considered as tentative until a more in-depth analysis of the data source is accomplished. The potential uses of these datasets for MOMENTUM are the analysis of the temporal variation of the demand, the distribution of distances travelled with these transport modes or the distribution of its use in the different car-sharing stations in the case of Regensburg. Furthermore, in this case study, as there are three years of data available, it will allow analysing how previous factors have evolved over time, how car-sharing demand may be influence by exogenous factors as weather conditions, holiday periods, etc. However, since we do not have information about the users, it will not be possible to infer the activity associated with the trips. Furthermore, it will limit the study of the factors influencing car-sharing e-scooter sharing adoption in Regensburg and Thessaloniki, respectively.

5.9 Traffic Data

5.9.1 Introduction

In the **Madrid Case Study**, the traffic data comes from the City Council, and it is also publicly available in their open data portal[1]. The traffic data is collected through more than 4000 measurement points, from which 253 are double loop sensors able to measure speed and characterize the vehicles. The measurement points are categorized in two classes: URB, for sensors that measure urban traffic and that are used for traffic light control, and M30, for sensors that measure interurban traffic in expressways and accesses to Madrid. Information about

[1] https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=

the location of the measurement points and their GPS coordinates are also available[1]. Historical data from July 2013 measured every 15 minutes is provided in ZIP, CSV and XLS format and updated monthly. For each measurement contains the next information[2]: sensor identification, data and hour, sensor category (URB, M30), traffic flow in vehicles/hour, occupancy in percentage, vehicle load (value in the range 0-100), average speed in kilometres per hour, an indicator of erroneous measurement and number of measurements considered for the integration period.

For the **Regensburg Case Study**, there are two data sources with traffic information. The first one it is provided by the State of Bavaria and its access is private. The data corresponds to traffic collected by sensors deployed in the road network run by the Bavarian Government. For the MOMENTUM project, only the four sensors located at Regensburg's Ring will be considered (sensor ids 9049, 9044, 9045 and 9009). The temporal scope of the sample provided is August 2018 to July 2019, and the temporal granularity is hourly. The data source contains the next information: date, day of the week, the hour of the day, total flow of vehicles, total flow of trucks and total flow in each direction for all vehicles and for trucks. The second one corresponds to the floating car data. TUM has started evaluating the FCD data from INRIX, as a representative. The data sample being evaluated has a temporal scope of just one day, and TUM is the only partner with access to this data. They are also exploring various other options available.

In **Leuven,** there are available two data sources for traffic. The first one is Telraam, was already described in Section 5.4. However, it is important to mention here for motorized traffic these devices are able to differentiate between car and heavy vehicles. The second one is provided by the Agentschap Wegen en Verkeer, the Belgian Agency for Roads and Transport. The data captures the number of cars, trucks, bicycles and pedestrians during the morning peak and the evening peak for 2 hours and queue lengths in 35 intersections with signal control in Leuven. The data was collected from 7 am to 9 am (morning peak) and from 4 pm to 6 pm (evening peak) in different days in 2018 and 2019 (one day per intersection).

Finally, for the **Thessaloniki Case Study**, there is one traffic data source that comes from the collaboration between Taxiway and CERTH-HIT mention in Section 4.3. The dataset consists of records with a timestamp, location and instantaneous speed (magnitude and orientation) of taxis that circulate in the broader area of Thessaloniki, Greece. This dataset's records are spatiotemporal vehicle pulses that utilize the Global Navigation Satellite System. In this case, the data attributes (coordinates, instantaneous speeds, etc.) are generated by the sensor of tablets which are planted in the taxis. The location's accuracy is typically within the order of a few meters. A new entry is typically generated for each active vehicle every 6 to 10 seconds. The data are being captured since 2013, and approximately 50 million entries are added each month. The data is available through FTP for historical datasets[3] (2017-2020) and Web Service for real-time access[4]. Historical data is updated monthly while the real-time data is refreshed every 15 minutes. Regarding data format, historical is a tab-delimited text file, whereas the real-time web service provides the information is JSON, XML and CSV. The data source contains the next fields: timestamp of the sample, longitude, latitude, altitude, speed and orientation (degree as an angle of speed from the North direction).

1 https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=ee941ce6ba6d3410VgnVCM1000000b205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD

2 https://datos.madrid.es/FWProjects/egob/Catalogo/Transporte/Trafico/ficheros/Estructura_DS_Contenido_Trafico_Historico.pdf

3 http://opendata.imet.gr/dataset/fcd-gps-historical

4 http://opendata.imet.gr/dataset/fcd-gps

5.9.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg		Leuven		Thessalon.
Type		Loop sensors	Loop sensors	Floating car data (e.g. INRIX)	Counts from Crowd- sourced Camera sensors	Intersection counts	Floating car data (Taxiway)
Reliability		Medium	Medium	High	High	High	High
Temporal & Geographical Scope	Last update	January 2020	July 2019	N/A	2019	2019	Continuous update
	Temp scope	2013-2020	Aug 2018 - Jul 2019	One day	2019	Specific days 7am-9am + 16pm-18pm	2013-2020
	Geo scope	City	Regensburg's Ring (4 sensors)	City	City	City	City
Gran. & Level of Detail	Temp gran	15 minutes	Hourly	Minutes	Hourly (only on daytime)	15 minutes	15 minutes
	Geo gran	At sensor level	At sensor level	Road segment	Road segment	Turns	Exact coordinates
	Agg. Lev	Aggregated measures	Aggregated flow per vehicle type	GPS pulses	Aggregated counts	Aggregated counts	GPS pulses
Availability & Accessibility		Whole project CSV/Excel Public	Whole project Excel All partners	Whole project CSV Specific partners through private agreement	Whole project JSON/CSV All partners	Whole project Excel All partners	Whole project JSON, XML, CSV Public(CC Non - Commercial license)
Relevance/Usability		High	Low	Low	Medium	Medium	High

Table 16. General overview of the Data Quality Assessment of Traffic Data

5.9.3 Identified issues

The main issues identified for the traffic data source in **Madrid** come from the problems inherent to the sensor technology that is inductive loop. Some of the most known problems are the trend to undercount vehicles (accuracy is usually around 95%), the number of lanes may influence detector errors, traffic composition (particularly heavy vehicles) may lead to overcount of vehicles, poor detection of small vehicles, undercount of vehicles in traffic lights intersections or high failure rates[1]. In this sense, the EMT has also reported the possibility of having many count locations with empty data in several years.

Regarding the **Regensburg Case Study**, the first data source listed shares the same issues as the previous one, as vehicle detections also come from inductive-loop sensors. However, the main drawback, in this case, is the very low number of sensors (only four). Apart from this, Regensburg also reported that the data provided by the metering point 9074 might be subject to errors due to the current construction site from 28.07.2019, which led to the lack of heavy vehicle counts. As for the INRIX trips and floating car data, the main issue that presents come from the fact that INRIX monitored vehicles represent only a small portion of actual vehicles in Regensburg, and it may be biased towards specific categories of fleets. Furthermore, it is important to point out that the sample corresponds to only one day of data, so there is a high risk that the analysis performed is biased by the special conditions of the reference day (e.g. weather, season, holidays, etc.).

In the **Leuven Case Study**, the Telraam data source present the same issues described in Sections 5.4.3 and 5.5.3, with the exception that for motorized traffic the detection accuracy of this devices is high and that the problem of camera orientation disappears as it only affects the sidewalks and not the road. As for the data source related to counts at intersections, its main drawback is the temporal scope, as it involves measures on specific days at certain intersections, and therefore subject to a high risk of bias. In any case, it is typical for this type of measurements given its high cost.

Finally, the traffic data in **Thessaloniki** presents the same issues defined for INRIX as it also comes from floating car data, but with some particularities. The vehicle sample providing data is smaller, and the bias towards specific categories of fleets is higher in this case as the monitored vehicle are taxis which have remarkably different travel and driving behaviour than the average user. However, in this case, the temporal scope is not a problem as the data is being recorded by CERTH since 2013. CERTH also reported that inaccuracies in GPS coordinates may be in the order of a few meters in roads with high buildings and obstacles, but so far, the GPS position, speeds and orientation accuracies have never caused significant problems in analysing the data.

5.9.4 Relevance/Usability

For the **Madrid Case Study**, the relevance of the available traffic data source is high. The data is publicly available with a good temporal and geographical scope and a good granularity with traffic measured every 15 minutes. The main potential use of this data source within MOMENTUM is the analyses of the demand for motorized road traffic. Although it is usual that traffic data coming from sensor loops allow differentiating between different categories of vehicles, in this case, it is not possible as all measures provided refer to averages over all types of vehicles. Given the large temporal scope of the data source, the influence of exogenous factors (e.g. weather, calendar, etc.) can also be studied. Other potential uses of this data source are the calibration of traffic models or OD matrices.

As the data collected by Telraam crowd-sourced devices in **Leuven** shares similar characteristics with inductive loops (e.g. they provide traffic counts at specific locations), the potential uses of this data source are the same. However, the relevance of this data source is medium since, as mentioned above, the deployment of these sensors

[1] Traffic Detector Handbook: Third Edition—Volume II, 2006.

in Leuven is irregular and leave some parts of the city undercovered, and the temporal scope is restricted to 2019 which limits historical analyses and the scope of the study of the influence of exogenous factors.

Regarding the manual counts at intersections in Leuven, its relevance is low since they are just punctual measures at specific intersections, as discussed above. The only potential use of this data source in MOMENTUM is the calibration of traffic micro-simulation models.

Turning to the case study in **Regensburg**, the relevance of the traffic data from the four loop sensors installed in this city's ring is low, mainly because of its reduced geographical coverage. The possible uses within MOMENTUM are very restricted and are practically limited to estimates of peak hours of traffic entering and leaving the city. The other available data source, floating car data, is more relevant (medium). Its main application within MOMENTUM is the estimation of travel times in Regensburg. Other potential uses would be the estimation of the traffic conditions in the transport network of this city or the creation of speed profiles in those road segments where there are a sufficient number of trajectories. However, the latter two uses will be limited by the small temporal scope of the available data sample.

As for the traffic data source available in the **Thessaloniki Case Study**, the potential uses are the same as those described for INRIX, although with fewer limitations as the temporal scope of this data source is extensive.

5.10 Taxi Service Demand Data

5.10.1 Introduction

Thessaloniki is the only use case with a data source available in this category. These are data concerning taxi trips in this city which also come from the collaboration between CERTH-HIT and Taxiway described in Section 4.3, and are therefore the property of Taxiway. As in the previous cases, CERTH already has access to this data source, and specific partners will have access to it through the signing of an NDA. These dataset's records are produced after processing the spatiotemporal taxi vehicle pulses recorded by onboard receivers of the Global Navigation Satellite System. All information for each realized trip from 2016 onwards is recorded, and the dataset is updated monthly. For each taxi trip recorded in the system, the data provided is the following: timestamps and GPS coordinates of the trip start and end locations, a flag indicating that the trip occurred with a passenger, trip duration in seconds and trip distance in meters. The dataset will be provided in a CSV file.

5.10.2 General overview of the Data Quality Assessment

Type & Data quality dim.	Madrid	Regensburg	Leuven	Thessaloniki
Type	N/A	N/A	N/A	Taxi trips
Reliability	N/A	N/A	N/A	Medium/High

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Temporal & Geographical Scope	Last update	N/A	N/A	N/A	Continuous update
	Temp scope	N/A	N/A	N/A	2016-2020
	Geo scope	N/A	N/A	N/A	City and suburbs/exurbs
Gran. & Level of Detail	Temp gran	N/A	N/A	N/A	Trip start/end event
	Geo gran	N/A	N/A	N/A	Exact coordinates
	Level of Detail	N/A	N/A	N/A	Trip
Availability & Accessibility		N/A	N/A	N/A	Whole project CSV Specific partners under NDA
Relevance/Usability					High

Table 17. General overview of the Data Quality Assessment of Taxi Service Demand Data

5.10.3 Identified issues

The data source available does not present relevant issues. The only minor concerns are the lack of traceable information about the vehicle or driver, but CERTH has access to it, and they reported that it might be available to other partners in the future. Furthermore, there is no information about the number of occupants of the taxi, just whether the cab is occupied or not.

Another issue might be the representativeness of the taxies belonging to the Taxiway association. Data may be biased towards taxis operating for this particular association. However, the Taxiway association represents about half of the cabs of Thessaloniki, and it is broad and plural enough to cover the various taxi services available at this region.

5.10.4 Relevance/Usability

The data source has high relevance for analysing taxi demand in **Thessaloniki** as there is complete information at the trip level, and it presents an excellent geographical and temporal scope. The potential uses of this data source in MOMENTUN are the study of different aspects related to the use of taxi as the temporal variation of the demand or the distribution of trip distances and times, travelled by taxi trips, etc. Because of the excellent temporal scope of this data source, it is also possible to study how these factors have evolved over this time or how they are affected by exogenous factors like weather or calendar. Furthermore, complementing this data source with land use and socio-demographic data it will be possible to estimate the activity patterns associated with the trips, and the factors that influence the adoption of this transport service.

5.11 Social Media Data

5.11.1 Introduction

As in previous demand data source category, **Thessaloniki** is the only case study with social media data available. The dataset available consists of geo-tagged events at social media platforms, for this city. This dataset's records are produced every 20 minutes, from 2016 onwards. An algorithm requests non-personal, geo-tagged data from Facebook's Developer API and processes the data to retrieve the number of check-ins per location of interest for the study area. The raw geo-tagged activity data are processed by CERTH. The API requests are posted sequentially for areas of the city, with a dynamic temporal interval so that it is ensured that the API request limit of service is not passed. The information collected in each record is the next one: the category of the geotagged location, the name of the geotagged location, the GPS coordinates of the location, the number of geo-tagged events in the location and the timestamp for which the activity is counted. This data source belongs to CERTH, and it will be provided to specific partners under an NDA agreement. The dataset will be distributed in CSV format.

5.11.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		N/A	N/A	N/A	Geo-tagged check-in event
Reliability		N/A	N/A	N/A	High
Temporal & Geographical Scope	Last update	N/A	N/A	N/A	Continuous update
	Temp scope	N/A	N/A	N/A	Since 2016
	Geo scope	N/A	N/A	N/A	City
Gran. & Level of Detail	Temp gran	N/A	N/A	N/A	20 minutes
	Geo gran	N/A	N/A	N/A	Venue location
	Level of Detail	N/A	N/A	N/A	Aggregated visit counts

Type & Data quality dim.	Madrid	Regensburg	Leuven	Thessaloniki
Availability & Accessibility	N/A	N/A	N/A	Whole project CSV Specific partners under NDA
Relevance/Usability	N/A	N/A	N/A	Medium/High

Table 18. General overview of the Data Quality Assessment of Social Media Data

5.11.3 Identified issues

The main issues identified for this data source are those common to social media data. First, a probable bias in the sample, since social media data usually present biases towards social media users (which are not representative of the whole population) and also towards discretionary and leisure activities. On the other hand, the dataset contains a high number of different venue categories (878). In many cases, some of these categories refer roughly to the same type of venue (e.g. Coffee Shop, Cafe, Cafeteria). This will require computational and manual work to reduce current classes to more meaningful ones. Finally, because of the Facebook privacy policy, the data correspond to the aggregated number of check-in events per venue location, so no traceability of the user is possible.

5.11.4 Relevance/Usability

The relevance of this data source for MOMENTUM is medium-high. Given that it has a good temporal and geographical scope and granularity, and it covers a wide range of categories of venues in Thessaloniki, the potential uses for MOMENTUM are the analyse of the activity patterns in this city. If it is combined with other demand data sources, especially those with a high geographical and temporal granularity, it is possible to study its relations with the demand for specific transport modes. Given the demand data sources available in Thessaloniki, taxi and bike-sharing would be the best candidates. However, as we only have aggregated information, only correlation analyses between mobility patterns and the number of visitors/check-ins in specific categories of venues can be performed.

5.12 Parking Data Demand

5.12.1 Introduction

In the **Madrid Case Study,** the parking demand data is provided by the EMT and comes from the fourteen public off-street parking run by this company. The dataset corresponds to snapshots of the occupancy of these parking facilities every three minutes, recorded since 2016. The access to the historical dataset is private, but it will be available for all partners during the whole duration of the project. There exist another complementary data source that consists of a public REST API[1] that provides information about parking availability in real-time together with other information as the postal address, a flag indicating whether the parking belongs to the EMT or not and the GPS coordinate of the parking. Regarding the historical dataset, it is provided in CSV format and for each recorded snapshot contains the following information: parking identifier, parking name, timestamp of the snapshot and the

[1] https://apidocs.emtmadrid.es/#api-Block_5_PARKINGS-parking_availability

number of free parking places at that time. Apart from this dataset, the next weeks, we will also explore the possibility of obtaining parking demand data from those facilities run by Madrid City Council.

The other case study with data about parking demand is **Leuven**. The data provider, in this case, is the City of Leuven and it contains information about the use of three off-street parking spaces in Leuven, concretely, from Minckeler, Ladeuze and Heilig Hart parking garages. No data from other parking were provided because of their heterogeneity in data collection procedure. Still, if required for project purposes, the demand data from other parking places will be available for the project. As for the data from the three off-street parking considered, the temporal scope of the sample provided rages from January 1st 2018 till 30th April 2019, although it is expected to be updated along 2020. The dataset is provided in Excel format, and it contains one record for each in-out event, that is, for every instant a vehicle enters or leave the parking. Concretely, each of this records has the following information: parking name, a flag indicating whether the parking is open or close, total number of free parking places, occupation percentage, total number of occupied parking places, percentage of occupied parking places, data and time of the in/out event. At the moment of writing this deliverable, Leuven also commented about the possibility of having demand information from on-street parking from May 2020. The data collection will be done for the whole territory of Leuven, including the outer city districts. They intend to map the parking spaces on 420 km of Leuven roads by using mobile mapping. Concretely, 360° images and point clouds will be collected by three vehicles and then used for making an inventory and analyzing the parking demand.

No information within these categories is available for **Regensburg** and **Thessaloniki** case studies. However, at the moment of writing this deliverable, in Thessaloniki we will explore the possibility of obtaining parking demand data through a data source that CERTH-HIT have been collected about parking status and driver's behavior on major roads of the city of Thessaloniki.

5.12.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Periodical snapshots of parking occupancy		Vehicle entry/ exit events	
Reliability		High	N/A	High	N/A
Temporal & Geographical Scope	Last update	2019	N/A	2019	N/A
	Temp scope	2016-2019	N/A	2018-April 2019	N/A
	Geo scope	City (14 parkings)	N/A	City (3 parkings)	N/A
Gran. & Level of Detail	Temp gran	3 minutes	N/A	Entry/Exit event	N/A
	Geo gran	At parking level	N/A	At parking level	N/A
	Level of Detail	Occupancy at snapshot timestamp	N/A	Aggregated number of entries/exits	N/A

Type & Data quality dim.	Madrid	Regensburg	Leuven	Thessaloniki
Availability & Accessibility	Whole project CSV All partners	N/A	Whole project Excel All partners	N/A
Relevance/Usability	Medium/High		Medium/Low	

Table 19. General overview of the Data Quality Assessment of Parking Data Demand

5.12.3 Identified issues

In the case study of **Madrid**, the main shortcoming of the data source is that it only covers the public parking places run by the EMT and no data is available about the usage of other private, public and semi-public parking facilities in Madrid, so the view on the use of off-street parking is partial. Furthermore, the recorded snapshots only contain information about the parking occupancy, and not on the incoming/outcoming flows.

For the **Leuven**, the main issue of the off-street parking information is the low number of parking facilities (only four) with appropriate and comparable data available. The analysis of data from other parking facilities would require an important effort as they have different data formats which make analysis and processing difficult. Besides, the mechanism for counting off-street parking spaces can be very different, so to obtain comparative data, further data processing is required.

5.12.4 Relevance/Usability

Madrid's data source is of medium-high relevance for MOMENTUM. Although it only provides a partial view of car park demand in Madrid (those car parks belonging to the EMT) and occupation information, the fact that this information is available every 3 minutes and for four years makes it a very interesting dataset. The potential uses within MOMENTUM are the analyse of general and local (at parking level) patterns and temporal variations in car park demand in Madrid, its relation with other transport modes through the joint analysis with other data source (e.g. bus smart card data to study the use of dissuasive car parks) and the influence of exogenous variables as weather, pollution episodes, etc.

Regarding the **Leuven Case Study**, the relevance of the parking demand data medium/low. The main reason is the availability of only three off-street parking places with appropriate information about incoming/outcoming flows since the others would require important efforts to obtain meaningful information. The potential uses of this data source within MOMENTUM are similar to those described for Madrid but with the limitations of only having information from three parking lots. Joint analysis with other sources of demand data for different modes of transport is very limited in this case as these have a high level of aggregation (e.g. bike-sharing, traffic).

6 Data Quality Assessment for Maps & Cartography Data Sources

6.1 Land Use Data

6.1.1 Introduction

In the **Madrid Case Study**, the land use dataset comes from the Spanish Geographic Information Centre (Centro Nacional de Información Geográfica - CNIG). It is publicly available, and it can be downloaded from CNIG downloads centre[1] (it is necessary to complete a survey to download the information). Land information is presented in SHP and GDB formats. The information provided is divided into homogeneous polygons associated to a description of its land use, from a range of categories defined by the responsible authority of the dataset[2] as building (differentiated by various types), artificial green area, roads, unbuilt land, crops, forest trees, etc. This dataset has been updated in four periods of time: 2005, 2009, 2011 and 2014.

For **Regensburg Case Study**, the land use data were obtained from the governmental planning tool, a graphic plan of the entire municipal area, in which the existing and the future desirable land uses are shown. It is a private data source that belongs to the City Council, but it will be available for all MOMENTUM partners during the whole duration of the project. The dataset is provided in SHP and DWG formats. The geographical scope of the data source is the City of Regensburg, and the geographical granularity is similar to that of Madrid. The information is provided in the form of homogeneous polygons (representing developmental areas) with their associated land use category that includes residential, business, mixed, industrial, road, educational, hospital, etc. The only available version of this dataset is from 2017, that is the most current one.

The data source available for the **Leuven Case Study** is the City GIS, a complete data source for land use with information at the building block level, that it is currently used for all kinds of planning and analytic purposes. The base of the GIS is the official large-scale reference map provided by the regional government, and it contains very detailed transport network data, including traffic lights, traffic signs and all potentially relevant objects in the public domain; land use data, based on categories of buildings and infrastructure; and information about points of interest that can be obtained from building types. The dataset was provided in SHP format, and it is accessible by all partners only for project purposes. The sample provided is a snapshot of the state of the GIS database on October 1st, 2019.

For the **Thessaloniki Case Study**, the data source consists of geospatial data describing the land use observed in the Municipality of Thessaloniki. The dataset is 1 of the 13 thematic layers of the Spatial Data Infrastructure (webGIS) of the Municipality of Thessaloniki[3], named Urban Planning, that it is publicly available and which contains the sub-layer, Zone Use and more sub-layers, such as neighbourhoods of the Municipality of Thessaloniki, new parcel boundaries, urban planning plans, building blocks ESYE, building limit line and more. Any person has easy and free access in the webGIS and can download the available data from a list of formats. The format choices are SHP, CSV, KML, TIFF, and PDF. In case SHP format is chosen, there is an option to download the file in the desired reference system (WGS 84, GGRS 87 etc.). We could not find information about the last update of this data source, but CERTH reported that the reliability is high. The geographical granularity, in this case, is at the building block level.

[1] http://centrodedescargas.cnig.es/CentroDescargas/buscadorCatalogo.do?codFamilia=SIOSE#
[2] https://www.siose.es/SIOSEtheme-theme/documentos/pdf/Estruc_Cons_Bas_dat_SIOSE_v3.pdf
[3] https://gis.thessaloniki.gr/sdi/

6.1.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		National Land Cover GIS Database	City GIS Database	City GIS Database	City GIS Database
Reliability		Medium	High	High	High
Temporal & Geographical Scope	Last update	2014	2017	October 2019	No Information
	Temp scope	2005, 2009, 2011, 2014	2017	2019	No Information
	Geo scope	Country	City	City	City
Gran. & Level of Detail	Temp gran	N/A	N/A	N/A	N/A
	Geo gran	Homogeneous land use polygon	Homogeneous land use polygon	Building Block	Building Block
	Level of detail	Land use category	Land use category	Land use category	Land use category
Availability & Accessibility		Whole project SHP, GDB, GPKB Public	Whole project SHP and DWG All partners	Whole project SHP All partners	Whole project CSV/KML/PDF/SVG/TIFF/ESRI SHP Public
Relevance/Usability		Medium/High	Medium/High	High	High

Table 20. General overview of the Data Quality Assessment of Land Use Data

6.1.3 Identified issues

Regarding the **Madrid Case Study**, the issues identified are the following. First, although the general documentation specifies common criteria, it must be considered that the dataset is prepared by each Region's officers, which may lead to slight differences in the classification of the same land use. However, as we will only work with information from the Madrid region, this problem is not relevant. Second, the original data include some geometric errors that may turn problematic when processing it. However, Nommon maintains a pre-processed database fixing these errors that will be provided to the consortium. And third, the last update of this dataset is 2014, and hence, some of the land uses may be outdated

In **Regensburg case study,** the data source most recent update was in 2017, which may lead to some obsolete information but taking into account that it is referred to land use it is not a relevant issue.

Regarding the data source available for **Thessaloniki,** the only issue found is the lack of information about the last update of the dataset. However, as we mention above, CERTH reported that the reliability of the data is high, so in principle, this aspect does not present a relevant problem.

No relevant issues have been identified for the **Leuven's** data source.

6.1.4 Relevance/Usability

The relevance of the **Madrid Case Study**'s data source for MOMENTUM is medium-high. It provides information about land cover and land use with good geographical granularity and scope, although the fact that it has not been updated since 2014 reduces its reliability. The potential uses of this data source in MOMENTUM are, on the one hand, the combination with mobile phone records (described in Section 5.7) to increase the spatial granularity of the indicators that can be extracted from them. And on the other hand, the joint analysis with some demand data sources with high spatial granularity and user traceability (e.g. mobile phone records, PT smart card and bike-sharing data) to characterize activities from the recorded locations from the land use observed in the area where the activity is registered.

In the case of **Regensburg**, **Leuven and Thessaloniki**, the relevance of the data sources available is high since they do not present important issues. Their potential use in MOMENTUM is analogous to the second one pointed out for Madrid. The combined use with mobile phone records in these cases studies is not possible because there is no such type of data source available in these scenarios, as we discussed in Section 5.7.

6.2 Weather Data

6.2.1 Introduction

In the **Madrid Case Study,** the weather data come from the Spanish Weather Agency (AEMET), and it is publicly available in its open data portal[1]. The meteorological data is obtained from official weather stations distributed all over Spain (800 weather stations) and are accessible on a real-time basis through a public REST web service. AEMET also maintains historical records of these data, but access to them is not free. For this reason, Nommon developed a Python script that collects the updates of the data hourly and gathers it in sorted files, producing one JSON file per day. The files contain one single line per hour record of each station along the day. The records may include the following fields: station identifier/common name, Latitude/Longitude, Timestamp, Liquid/Solid precipitation, Temperature, Pressure, Wind speed/direction, Humidity, Sunshine duration, etc. Not all stations can produce all kind of data; thus, some of these fields are optional. This weather database started to be collected in April 2019, and it will be accessible to all MOMENTUM partners.

For the city of **Regensburg**, no official weather data source was identified with free available weather data. For this reason, we had to resort to web services with historical weather information. Two of the best known and most used are Darky.net[2] and Weatherunderground[3]. The first service bases its meteorological data on the aggregation of different data sources (the list of sources can be consulted on its website[4]), while the second is based on data provided by amateur weather stations voluntarily deployed by users worldwide, which makes it slightly less reliable. In both cases, the historical information is quite rich, at local level, and with hourly values on temperature, rainfall, humidity and wind, among others. In the case of DarkSky.net, access to the historical data is done through a REST API. For Weatherunderground it has to be done through a web form. These services have historical records for Regensburg from 2016 and from 2015, respectively.

[1] https://opendata.aemet.es/centrodedescargas/inicio
[2] https://darksky.net/dev
[3] https://www.wunderground.com/
[4] https://darksky.net/dev/docs/sources

Regarding the city of **Leuven**, the official weather data source available is the Royal Meteorological Institute of Belgium (KMI) which have about 20 weather stations spread over whole Belgium. The closest to Leuven is located in Uccle, around 25 km far from Leuven. It provides hourly historical data from 2015 with the following information: average, high and low temperature in ºC; average, high and low dewpoint in ºC; average, high and low humidity; max and min pressure in hPa; max and average wind speed; max gust speed; and total precipitation in cm. The dataset is open, and it can be obtained in Excel format from the National Centers for Environmental Information[1]. Given the low availability of weather stations close to Leuven from the KMI agency, the two weather data services pointed out for the Regensburg case study will also be considered here.

In the **Thessaloniki Case Study,** the dataset was obtained from the online platform openweathermap.com[2]. CERTH-HIT retrieved historical weather data and initiated the archiving of real-time weather from 2012 onwards. The dataset consists of records about the current conditions in the city of Thessaloniki. The fields describe several weather characteristics and related measurements, including temperature, pressure, humidity, wind speed and direction, cloud coverage and a generic, qualitative weather description (e.g. clear sky). The dataset is stored in JSON format with a minute granularity, and it will be only available for some partners under an NDA.

6.2.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		National Weather Agency	DarkSky.net/Weatherundergound	National Weather Agency	Openweathermap.com
Reliability		High	Medium/High	Medium	High
Temporal & Geographical Scope	Last update	Continuous update	Continuous update	Continuous update	Continuous update
	Temp scope	Since April 2019	From 2015	From 2015	From 2012
	Geo scope	Country	Worldwide	City (One weather station)	City
Gran. & Level of Detail	Temp gran	Hourly	Hourly	Hourly	Minutely
	Geo gran	Weather station	Weather station	Weather station	Weather station
	Level of Detail	Detailed weather conditions	Detailed weather conditions	Detailed weather conditions	Detailed weather conditions

[1] https://www.ncdc.noaa.gov/
[2] https://openweathermap.org/

Type & Data quality dim.	Madrid	Regensburg	Leuven	Thessaloniki
Availability & Accessibility	Whole project JSON All partners	Whole project JSON, Excel Public	Whole project CSV, Excel, PDF Public	Whole project JSON Specific partners (NDA)
Relevance/ Usability	High	High	Medium	High

Table 21. General overview of the Data Quality Assessment of Weather Data

6.2.3 Identified issues

Regarding the **Madrid Case Study**, the only issue found is the temporal scope of the dataset available since data started to be collected in April 2019. If necessary, this limitation may be solved by using other services are DarkSky.net or Weatherunderground.

In the case of **Regensburg,** as there is no official data source, the most convenient options are general weather services as weatherunderground.com or DarkSky.net as explained above. As for the first one, the only limitation identified is that the weather data come from amateur weather stations which may lead to less accurate measurements, although some reports confirm medium/high reliability. For DarkSky.com, the only issue identified is the limitation of 1000 free data queries per day. Given that we will work with data at the region level, this is not a relevant limitation, and the cost of extra data queries is low (0,0001$ per data query).

The main shortcoming for the data source available in **Leuven** is that the data is only provided by one weather station at 25 km from Leuven, which could lead to inaccurate weather condition measurements. This issue can be solved by using general weather information services like the one mentioned in the previous paragraph.

No relevant limitations were observed for **Thessaloniki**'s dataset.

6.2.4 Relevance/Usability

Given that, in all case studies, we can have available weather data sources with medium/high or high reliability, with at least hourly temporal granularity, with weather station or exact location geographical granularity, and with a reasonable temporal scope, the relevance in all cases can be considered high. The potential uses of these data sources in MOMENTUM are, on the one hand, to understand how bad weather conditions affect the use of certain emerging mobility services (e.g. bike-sharing, car-sharing, etc.). And, on the other hand, to analyse how the users of those transport modes that could be more sensitive to adverse weather conditions (e.g. cyclists, pedestrians, etc.) react to these conditions in cities with different climatology.

6.3 Social, Cultural or Sportive Events

6.3.1 Introduction

The only case study with information available in this subcategory is **Leuven.** Concretely, the data source is the Uitdatabank[1]. Uitdatabank is a database of events in Flanders, where every year, more than 28,000 organizers enter more than 215,000 activities for free. Events are submitted by users and checked by administrators per

[1] https://www.uitdatabank.be/

region. The database can then be used to obtain event agendas on countless websites via a (paid) API[1]. Data can be obtained in JSON format, and it is composed of the following fields: event name, location, start and end date, labels (e.g. event type and theme). The Uitdatabank database keeps records since 2006, and it is updated every day.

6.3.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		N/A	N/A	Crowd-sourced event database	N/A
Reliability		N/A	N/A	High	N/A
Temporal & Geographical Scope	Last update	N/A	N/A	Continuous update	N/A
	Temp scope	N/A	N/A	From 2016	N/A
	Geo scope	N/A	N/A	City of Leuven	N/A
Gran. & Level of Detail	Temp gran	N/A	N/A	Minutely	N/A
	Geo gran	N/A	N/A	Exact locations	N/A
	Level of Detail	N/A	N/A	Dates, description and labels	N/A
Availability & Accessibility		N/A	N/A	Whole project JSON All partners	N/A
Relevance/Usability				High	

Table 22. General overview of the Data Quality Assessment of Social, Cultural or Sportive Events

6.3.3 Identified issues

The only issue identified is the difficulty to differentiate between small and large events (that are the most interesting for MOMENTUM) which must be done from the labels associated with the events.

6.3.4 Relevance/Usability

The relevance of this dataset is high for MOMENTUM because of its good geographical and temporal granularity and scope, as well as detailed information about events. The potential application of this data source within MOMENTUM is the analysis of the use of shared-mobility services for attending to large or medium events.

[1] https://documentatie.uitdatabank.be/

6.4 Points of Interest

6.4.1 Introduction

For this subcategory, there are no data sources available for **Madrid** and **Regensburg** case studies.

In the case of **Leuven**, the Points of Interest can be extracted for the GIS dataset described in Section 6.1, from building classification. In this way, the characteristics of this data source are analogous to the GIS dataset with land use information in Leuven.

In the **Thessaloniki Case Study**, the situation is similar. The data was extracted from the same data source as land use data for Thessaloniki, also described in Section 6.1. The data source contains information about conservancy of antiquities monuments, municipal services, parking spaces, public market boundaries, public market, public order, culture, religion, shops of sanitary interest, open care centre for the elderly, public areas, accommodation, kiosks, act of adjustment, public service, pre-nursery schools, nursery schools, primary schools, secondary schools, high schools, technical schools, health, public transport, taxi stations, playgrounds, municipal properties, disabled access friendly services, parking for people with disabilities etc.

6.4.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		N/A	N/A	City GIS Database	City GIS Database
Reliability		N/A	N/A	High	High
Temporal & Geographical Scope	Last update	N/A	N/A	October - 2019	No Information
	Temp scope	N/A	N/A	2019	No Information
	Geo scope	N/A	N/A	City of Leuven	Prefecture of Thessaloniki
Gran. & Level of Detail	Temp gran	N/A	N/A	N/A	N/A
	Geo gran	N/A	N/A	Building or POI	Building or POI
	Level of Detail	N/A	N/A	POI category	POI category
Availability & Accessibility		N/A	N/A	Whole project SHP All partners	Whole project CSV, KML, PDF, SHP All partners
Relevance/Usability				High	High

Table 23. General overview of the Data Quality Assessment of Points of Interest

6.4.3 Identified issues

The issues identified are analogous to the ones described in Section 6.1.3

6.4.4 Relevance/Usability

The relevance of the data sources available for Leuven and Thessaloniki is high because of their good geographical scope and granularity as well as reliability. The potential uses of these data sources within MOMENTUM are the inference of certain activity patterns associated to trips, although this requires a demand data source with high geographical granularity and user traceability, which at the moment of writing this deliverable, are available in Thessaloniki case study but not in Leuven.

7 Data Quality Assessment for Socio-Demographic Data Sources

7.1 Demographic statistics

7.1.1 Introduction

In the **Madrid Case Study**, the demographic statistics available are public and come from the Spanish Statistics Institute (Instituto Nacional de Estadística - INE)[1]. The population figures of the Register, with reference to January 1st of each year, are obtained according to registry regulations. The INE disseminates them once the Royal Decree of approval thereof has been published in the BOE. The data of the population resident in Spain is offered on January 1st of each year, according to the following classification variables: the place of residence, gender, age, nationality and place of birth. The place of residence data is provided for different levels of territorial disaggregation: national, autonomous communities, provinces, municipalities and census tracts. The temporal scope of this data source is 2008 - 2019, and they are updated yearly.

For **Regensburg Case Study**, the demographic data is provided by the City Council in the so-called "Statistical Yearbook"[2] that is publicly available. The Statistical Yearbook provides a compilation of information and figures aggregated for the City of Regensburg and each of the 18 urban districts. It is a complete report with statistics about demographics, building and housing, industry, labour market, supply and consumption, transport, tourism, healthcare, social services, education, culture and leisure, environment protection, public safety and order, and municipal finance and administration. Regarding demographic data, it provides information about the current population by age group, births and deaths, marriages and divorces per 1,000 inhabitants between 1966 and 2018. From 2000, the demographic statistics reported have been based exclusively on the data available at the end of each month in the population register. Before that year, the population statistics reported corresponds to the extrapolation of censuses. The report is available in PDF format, and the statistics can be downloaded in Excel format from the links embedded in the PDF document.

Regarding **Leuven Case Study**, the demographic data was obtained from Statbel (Belgian Statistical Office). The data are publicly available in the open-data portal of this institution[3]. According to the metadata available in this portal, it provides information on trends in the resident population figure as recorded on the National Register of Natural Persons (RNPP), defined by law. They contain essentially two types of measures: 1) measures relating to the population's characteristics at the beginning and middle of the year (i.e. January 1st and July 1st); and 2) measures related to population changes by various event categories (births, deaths, immigration, emigration, changes in marital status or nationality). The demographic data is available for each year since 2009 can be downloaded in Excel format with a geographical granularity at statistical sector level (approximately 1000 people).

In **Thessaloniki Case Study,** the dataset consists of records of demographic characteristics of the population residing in the Prefecture of Thessaloniki. Each one of the Municipalities of Thessaloniki corresponds to a unique code (ID), as well as, each building block within the respective municipality. Given this detail, in terms of geographical area, there are connected records with the above IDs, describing the Population (number of inhabitants), the Gender (Number of Males, Females) and the Age Category (number of residents less than 188 years old, 85 + years old etc.).

[1] https://www.ine.es/dynt3/metadatos/es/RespuestaDatos.html?oe=30260
[2] http://www.statistik.regensburg.de/publikationen/jahrbuch.php
[3] https://statbel.fgov.be/en/open-data

The data come from the 2011 Population and Housing Census that was a population census in Greece conducted by the Hellenic Statistical Authority on behalf of the Greek state between 10 and 24 May 2011. Regarding the geographical granularity of the data, CERTH has a mapped the building blocks to the Traffic Zones (TAZs) of the Prefecture of Thessaloniki. The (TAZs) represent areas of various levels of geographical combination (e.g. special use areas/buildings such as schools, industrial areas, Municipalities, etc.). They are objects that include traffic generation and attraction information. They are defined by their bounding polygons and their centroids with unique X and Y coordinates. The Prefecture of Thessaloniki consists of 329 TAZs, as 1 zone has been assigned to the neighbouring Prefectures, while the other Regions of Greece have been codified with greater integration.

7.1.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Population Register	Population Register	Population Register	Census
Reliability		Medium/High	High	High	High
Temporal & Geographical Scope	Last update	November 2019	2019	December 2019	2011
	Temp scope	2008-2019	2017-2019	2009-2019	1951 - today
	Geo scope	Country	City	City	Country
Gran. & Level of Detail	Temp gran	Yearly	Yearly	Yearly	Every 10 years
	Geo gran	Census tract (between 1000 and 2000 inhabitants)	City districts (between 992 and 28.260 inhabitants)	Statistical Sector	Building block
	Level of Detail	Aggregated Statistics	Aggregated Statistics	Aggregated Statistics	Aggregated Statistics
Availability & Accessibility		Whole project CSV, JSON, Excel Public	Whole project PDF/Excel Public	Whole project CSV/Excel, gpkg Public	Whole project Excel Specific partners (NDA)
Relevance/Usability		High	Medium	High	Medium

Table 24. General overview of the Data Quality Assessment of Census Data

7.1.3 Identified issues

Starting with the data source for the **Madrid Case Study**, the only issue found, although not relevant, is the medium/high reliability. The reason is the concerns about the floating population in several areas where the registered population may not be a realistic approach to the actual number of people living there.

Regarding the **Regensburg Case Study,** the available data source presents a relevant issue in the geographical granularity. Demographic data provided is disaggregated by urban districts, which can be considered a coarse-grain granularity for this type of data sources since there is a total of 18 districts. Furthermore, the distribution of the population among these districts is quite heterogeneous, ranging from 992 people up to 28.260 inhabitants, according to the last report with data from 2019.

In **Leuven,** the main limitation of the demographic dataset is that the registered population do not consider all students because, by Belgian law, students are registered in their parent's municipality. Taking into account that in Leuven, there are around 50.000 students, they represent around one-third of the actual population. To fill the gap, the number of students is estimated using data on student housing, although the reliability of this data source is intermediate.

In the case of **Thessaloniki,** the main issue of the demographic data source provided is that the data have not been updated since the 2011 census.

7.1.4 Relevance/Usability

The relevance of the data sources provided for demographic data taking into account the issues described above is high for Madrid and Leuven, and medium for Regensburg and Thessaloniki. The potential uses of these demographic data sources for MOMENTUM project are the expansion of samples by home location and the realization of demographic analyses.

7.2 Income statistics

7.2.1 Introduction

In the **Madrid Case Study**, the dataset comes from a collaboration agreement between the Spanish Statistics Institute, which provides census data, and Spanish Tax Agency, which includes income tax data and it is publicly available[1]. The data source consist of a file per indicator or group of indicators: average income per person or household (€), income source distribution (wage /retirement pension/unemployment benefit/other benefits/ other income) (%), population with income per consumption unit under 5k€, 7.5k€ and 10k€ thresholds (%), population with income per consumption unit under 40%, 50%, 60%, 140%, 150%, 160% and 200% of the median (%), average population age (years), population under 18 years old, population above 65 years old, average household size and single-person households (%). Additionally, some indicators present information segmented by gender, age groups, nationality, etc. In this case, the temporal scope of the available data is from 2015 to 2016, and the update was foreseen in December 2019 with income data from 2017. Furthermore, the data source update frequency is unknown, being its first publication in 2019.

As for the **Regensburg Case Study**, the dataset was also obtained from the "Statistical Yearbook" described in Section 7.1.1. The information provided in this case is the average household income per person (chapter 19.18) at the city level for 1991, 1995 and yearly form 2000 till 2017.

For **Leuven Case Study**, the income statistics were also obtained from Statbel's open-data portal. They correspond to fiscal statistics on income subject to personal income tax by statistical sector. It provides different statistics per

[1] https://www.ine.es/experimental/atlas/exp_atlas_tab.htm

statistical sector about, for example, total net taxable income, total net income global and per different taxation criteria (e.g. immovable and movable assets, professional activities, etc.). This dataset can be downloaded in Excel format, and it is updated yearly. The current temporal scope for this data source is 2005-2017.

In the **Thessaloniki Case Study**, the dataset available was obtained from the income data collected during the period 2017 – 2018, in the context of the implementation of the Origin-Destination Survey in households of the Prefecture of Thessaloniki described in Section 5.6.

The household income, as well as the personal income of the individual fulfilling the questionnaire, were included. The geographical granularity, in this case, is traffic zone (see Section 7.1.1 for description).

7.2.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Income Tax statistics from National Tax Agency	Average Household income statistics from Regional Government	Income tax data from National Tax Agency	Income statistics from Household Survey
Reliability		High	High	High	Medium
Temporal & Geographical Scope	Last update	November 2019	2019	December 2019	2018
	Temp scope	2015 - 2017	1995 - 2019	2005-2017	2018
	Geo scope	Country	City	Country	Prefecture of Thessaloniki
Gran. & Level of Detail	Temp gran	Yearly	Yearly	Yearly	Yearly
	Geo gran	Census tract (subdivisions of municipalities)	City	Statistical Sector	Traffic Zone
	Level of Detail	Aggregated Statistics	Aggregated Statistics	Aggregated Statistics	Aggregated Statistics
Availability & Accessibility		Whole project JSON, CSV, Excel All partners	Whole project PDF/Excel Public	Whole project CSV/Excel, gpkg Public	Whole project Excel Specific partners (NDA)
Relevance/Usability		High	Low	High	Medium

Table 25. General overview of the Data Quality Assessment of Income Statistics

7.2.3 Identified issues

In the data source available for the **Madrid Case Study**, the main issues found are, on the other hand, related to the conditions required for preserving statistical secret, which make the data totally or partially unavailable for some census tracts: the census tract must have more than 100 inhabitants, and for each indicator, the tail of the distribution among tracts is cut: those tracts with values under 0.5 percentile or above 99.5 percentile appear with the value for the respective percentile.

The data source collected for **Regensburg** presents a relevant limitation that is the low geographical granularity as we only have statistics at the city level.

Regarding **Thessaloniki Case Study**, it also presents a relevant issue. The data available come from a survey and not from tax income statistics, and therefore more prone to errors.

For Leuven, no shortcomings have been identified for the available data source.

7.2.4 Relevance/Usability

The relevance of the data sources for Madrid and Leuven is high because of their good geographical and temporal scope and granularity. In the case of Thessaloniki, the relevance is medium because of the reliability of the income statistics, as they should be derived from a survey. Finally, for Regensburg, the relevance is low because of the coarse-grained granularity, which limits its applicability. The potential uses of these data sources for MOMENTUM are to characterise mobility patterns in different income groups.

7.3 Tourism statistics

7.3.1 Introduction

In the **Madrid Case Study**, the tourism statistics come from the Hotel Occupancy Survey published by the INE (Spanish National Statistics Agency) and are publicly available in the City Council's webpage[1]. It provides monthly estimations for hotel beds, accommodation, distinction by categories, occupation by month, occupation by place of origin, occupation by category, etc. The temporal scope is 2019 and the different tourism variables considered are aggregated at the city level. The data can be downloaded in Excel format.

For **Regensburg Case Study**, the dataset was also obtained from the "Statistical Yearbook" described in Section 7.1.1. In this case, the information collected is operations, beds, arrivals and overnight stays, overnight stays by foreign visitors, arrivals by month, overnight stays by month and youth hostel occupation. The temporal scope is 2004-2018, and the geographical scope and granularity are the City of Regensburg.

In the **Leuven Case Study**, the data was again obtained from Statbel, and it is publicly available. It contains yearly information at the city level of total arrivals, total overnight stays, arrivals and stays for leisure and arrivals and stays for business. There is no data on visitors who stay with private persons. The data is available in CSV format, and the temporal scope is 2014-2018.

There is no tourism statistics data source available for **Thessaloniki.**

[1] https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Turismo-y-eventos/Turismo/Encuesta-de-Ocupacion-Hotelera/?vgnextfmt=default&vgnextoid=5ca8525b2a969210VgnVCM2000000c205a0aRCRD&vgnextchannel=6c3cf9b50632a210VgnVCM1000000b205a0aRCRD

7.3.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Survey	Survey	Survey	N/A
Reliability		High	High	High	N/A
Temporal & Geographical Scope	Last update	2018	2019	December 2019	N/A
	Temp scope	2018	2014-2018	2014-2018	N/A
	Geo scope	City	City	City	N/A
Gran. & Level of Detail	Temp gran	Monthly	Monthly	Yearly	N/A
	Geo gran	City	City	City	N/A
	Level of Detail	Aggregated statistics	Aggregated statistics	Aggregated statistics	N/A
Availability & Accessibility		Whole project Excel Public	Whole project PDF, Excel Public	Whole project CSV Public	N/A
Relevance/Usability		Low	Low	Low	N/A

Table 26. General overview of the Data Quality Assessment of Tourism Statistics

7.3.3 Identified issues

The data sources available for the cases studies in Madrid, Regensburg and Leuven, share the same relevant issue that is the coarse-grain geographical granularity of the data, concretely at the city level. Another important issue for the data source in Madrid is its partial view of the accommodation in the city, as it only considers hotel and no other type of accommodation as other apartments or platforms as AirBnB.

7.3.4 Relevance/Usability

Given the low geographical granularity of the tourism data sources available, the relevance of these data sources for MOMENTUM is low. No relevant potential uses for this project have been identified.

7.4 Car Ownership

7.4.1 Introduction

In the **Madrid Case Study**, the data source available was collected from the City Council's Vehicle Tax Data Base that is publicly available[1]. It provides different statistics about car ownership and car motorization rate as existing vehicle fleet by district and neighbourhood according to type of vehicle and fiscal power, existing vehicle fleet of natural persons by district and neighbourhood according to type of vehicle and fiscal power, existing vehicle fleet of legal entities by district and neighbourhood according to type of vehicle and fiscal power, existing vehicle fleet by vehicle type according to age, existing car park by district according to vehicle age, density indicators of the existing car park per district, existing vehicle fleet by census tract according to vehicle type and fiscal power, existing vehicle fleet of natural persons by census section according to type of vehicle and fiscal power and existing vehicle fleet of legal persons by census section according to type of vehicle and fiscal power. The temporal granularity of the statistics is yearly, and there are data available for the years 2016, 2017 and 2018. Regarding the geographical characteristics, the scope is the city of Madrid, and the granularity is up to census tracts. The data are available in Excel data format.

As for the city of **Regensburg**, the dataset was also obtained from the "Statistical Yearbook" described in Section 7.1.1 whose source is the German Federal Transport Authority. For this subcategory, the information available (Chapter 7.7) is existing vehicle fleet in total, for passenger cars (disaggregated by engine size), for motorbikes (disaggregated by type), for trucks (disaggregated by allowed maximum weight), for buses and coaches, for tractors, for trailers, and other types of motor vehicles. The temporal scope is 2014-2018; the geographical scope is the City of Regensburg; whereas the geographical granularity this time is at the district level.

For **Leuven Case Study**, the best data source available within this category is the car availability reported in the City Monitor Survey described in Section 5.6. Although the City Council keeps another data source with real registered cars, it is not very useful due to the high number of leasing cars registered in Leuven by large leasing companies based here for employees all over Flanders. The City Monitor Survey provides information about the number of vehicles and motorbikes per household, as well as the share of households with at least one car and the average number of cars per household. The geographical and temporal granularity and scope would be the same as the City Monitor Survey.

Concerning **Thessaloniki Case Study**, the data source is the 2011 Census already described in Section 7.1.1. The information is provided at the building block level and offers the following information related to car ownership: number of households with no cars, with one car, with two cars and with three cars.

[1] https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Trafico-transportes-y-comunicaciones/Parque-movil/Impuesto-sobre-vehiculos-de-traccion-mecanica-IVTM-/?vgnextfmt=default&vgnextoid=cc8f23a5e6c40510VgnVCM2000000c205a0aRCRD&vgnextchannel=9f51fbc78432a210VgnVCM1000000b205a0aRCRD

7.4.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Vehicle Tax Database	Vehicle Tax Database	Survey	Census
Reliability		Medium	High	Medium	High
Temporal & Geographical Scope	Last update	November 2019	December 2019	2019	2011
	Temp scope	2016 - 2018	1998 - 2019	2019	2011
	Geo scope	City	City	City	Country
Gran. & Level of Detail	Temp gran	Yearly	Yearly	Yearly	Yearly
	Geo gran	Census tract	Urban districts	Statistical Sector	Building block
	Level of Detail	Aggregated Measures	Aggregated Measures	Aggregated Measures	Aggregated Measures
Availability & Accessibility		Whole project Excel Public	Whole project PDF, Excel Public	Whole project Excel All partners	Whole project Excel All partners
Relevance/Usability		High	High	Medium	Medium

Table 27. General overview of the Data Quality Assessment of Car Ownership

7.4.3 Identified issues

Starting with the case study in **Madrid**, the issues found are the following:

- There are some outliers identified in the historical data (>500% growth in vehicles registered for certain areas from one year to another, followed by a return to average values in the next year)
- The segmentation depending on titularity can be used for a more accurate estimation of car ownership since total vehicle data includes those owned by companies and public bodies; natural person data ("personas físicas") corresponds to private vehicle owners.
- It can be seen that certain areas with a lot of companies or public bodies headquarters have a significant gap between total vehicle numbers and natural person vehicle numbers.
- The different tax policies of municipalities across Madrid region implies that not everybody registers the vehicle in the place of residence, but in municipalities with lower taxes and therefore, the car ownership metrics obtained from this dataset have to be interpreted as a low threshold.

The data source collected for **Regensburg** presents the same relevant limitation as in previous cases, the low geographical granularity as we only have statistics at the city level.

Regarding **Leuven Case Study**, it also presents a relevant issue. The data available come from a survey and not from a tax vehicle database, and therefore it is more prone to errors.

In the case of **Thessaloniki**, the main issue of the car ownership data source provided is the bad timeliness since the data corresponds to the 2011 census.

7.4.4 Relevance/Usability

The relevance of the car ownership data sources for Madrid, Regensburg, Leuven and Thessaloniki are Medium/High, Low, Medium and Medium, respectively, taking into account the issues described above. The most interesting potential use of these data sources within MOMENTUM is to identify relations between car ownership rates and use and adoption of shared mobility services. However, considering the relevance of the data sources available within this sub-category, the ones that can be used for this purpose are those from Madrid, Leuven and Thessaloniki.

7.5 Labour market statistics

7.5.1 Introduction

For the **Madrid** case study, labour market statistics can be obtained from the income statistics data source described in Section 7.2.1. This data source provides information about the income source distribution that includes wages and unemployment benefits, providing information about the employed and unemployed population.

In **Regensburg**, the dataset was obtained again from the "Statistical Yearbook" described in Section 7.1.1. In this case, the source is the Bavarian State Office for Statistics. For this subcategory, the information available (Chapter 5) are the employees subject to social insurance contributions; the employees subject to social insurance at the place of residence; as well as active population, employed and unemployed population among others. The temporal scope is 2014-2018, the geographical scope is the City of Regensburg, whereas the geographical granularity this time is at the urban district level.

Regarding **Leuven** Case Study, the dataset was collected from the Crossroads Bank for Social Security of Belgium (KSZ). The information available in this case is the sum of employed and unemployed persons, employed persons, unemployed persons and inactive persons. It is accessible publicly in KSZ webpage[1] through a web application and can be downloaded as an Excel file. The geographical scope is Belgium, and the statistics are provided at the statistics zone level. The temporal scope is very good as there are records from 2005, and the data are updated yearly.

In the case of **Thessaloniki**, the data source is again the 2011 Census already described in Section 7.1.1. This time, the dataset consists of records of labour market statistics in the Prefecture of Thessaloniki. In more detailed, the characteristics of Labour and Unemployment are presented at the level of Municipality, through the categories - Financial Active and Financial Non-active. The Financial Active population are the employed, unemployed, youths and previously employed. The Financial Non Active population are the students, elderlies and others.

7.5.2 General overview of the Data Quality Assessment

[1] https://www.ksz-bcss.fgov.be/nl/dwh/homepage/index.html

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Income tax data from the National Tax Agency	Regional Statistics Agency	National Crossroads Bank for Social Security	Population Census
Reliability		High	High	High	High
Temporal & Geographical Scope	Last update	November 2019	2019	December 2019	2011
	Temp scope	2015 - 2017	Since 2017	Since 2005	1951 - today
	Geo scope	Country	City	Country	Country
Gran. & Level of Detail	Temp gran	Yearly	Yearly	Yearly	Every 10 years
	Geo gran	Census tract	City districts	Statistic zone	Building Block
	Level of Detail	Aggregated Statistics	Aggregated Statistics	Aggregated Statistics	Aggregated Statistics
Availability & Accessibility		Whole project JSON, CSV, Excel Public	Whole project PDF, Excel All partners	Whole project Excel Public	Whole project Excel Specific partners (NDA)
Relevance/Usability		High	Medium/Low	High	Medium

Table 28. General overview of the Data Quality Assessment of Labour/Unemployment statistics

7.5.3 Identified issues

In the **Madrid and Thessaloniki** case studies, the issues found are the same as those discussed for the datasets on income statistics in Section 7.2.3 as they come from the same source.

As for the data source available for **Leuven,** we have not identified any relevant issue. For **Regensburg**, the main issue is again the coarse-granularity. However, unlike income statistics, we have some statistics (Employees subject to social insurance contributions, Unemployed) disaggregated by the urban district.

7.5.4 Relevance/Usability

The relevance of the data sources available for Madrid and Leuven is high because of their good characteristics in terms of geographical and temporal granularity and scope. Regarding Thessaloniki's data source, the relevance is medium (because of the bad timeliness), whereas this is medium/low for the one in Regensburg. The potential uses of these data sources within MOMENTUM are to characterise mobility patterns in groups of people with

different labour situations, and also to study the impact that labour market factors, such as the unemployment rate in an area, have on the adoption of emerging modes of mobility.

7.6 House price statistics

7.6.1 Introduction

Both **Madrid** and **Thessaloniki** do not have data sources available within this category.

For **Regensburg**, the datasets available within this category are two reports from the City Council and Regensburg Sparkasse, respectively. The first one is relative to 2018 and corresponds to the rent index elaborated by the City Council for the year 2018, and it is publicly available[1] in PDF format. The information is disaggregated by the urban district. The second one is a report from 2018 about the real estate and rental market in Regensburg, with information on housing and rental price evolution in the city and region of Regensburg. It is also publicly available in PDF format[2].

In **Leuven,** the available information comes again from Statbel, and it is related to the real estate sales according to the nature of the property in the deed of sale by statistical sectors. Concretely, for each statistical sector and house type, it provides yearly information about the number of sale transactions occurred and the percentiles 25, 50 and 75 of the sales prices. The data is publicly available and can be downloaded in Excel format[3]. Apart from this, Leuven provided the median house price for each statistical sector calculation from data of the last three years, averaged and scaled to the prices of 2019. For sectors which have nevertheless too few transactions, inverse distance weighted interpolation was used. The temporal scope of this data source is 2013-2018.

7.6.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type			Rental Index/Real State DataBase	Real State Sales Statistics	
Reliability		N/A	High	Medium	N/A
Temporal & Geographical Scope	Last update	N/A	2018	2019	N/A
	Temp scope	N/A	2009 - 2016	2013 - 2018	N/A
	Geo scope	N/A	City	City	N/A

[1] https://www.regensburg.de/fm/RBG_INTER1S_VM.a.253.de/r_upload/mietspiegel-2018-2019-stand-24012018.pdf
[2] https://www.sparkasse-regensburg.de/content/dam/myif/spk-regensburg/work/dokumente/regional/pdf/Sparkasse%20Immoreport%202018.pdf?n=true
[3] https://statbel.fgov.be/en/open-data/real-estate-sales-according-nature-property-deed-sale-statistical-sectors

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Gran. & Level of Detail	Temp gran	N/A	Yearly	Yearly	N/A
	Geo gran	N/A	City districts	Statistical Sector	N/A
	Level of Detail	N/A	Aggregated measures	Aggregated measures	N/A
Availability & Accessibility		N/A	Whole project PDF Public	Whole project CSV, Excel Public	N/A
Relevance/Usability		N/A	Low	High	N/A

Table 29. General overview of the Data Quality Assessment of House price statistics

7.6.3 Identified issues

Regarding the reports available in Regensburg, the main issue identified is the format of the information as both reports are only available in PDF, which makes difficult to extract the useful information, requiring an important effort.

As for Leuven's data source, no relevant issues have been identified.

7.6.4 Relevance/Usability

The relevance of the data sources available for Regensburg and Leuven is low and high, respectively, taking into account the issues above. The potential uses of these data sources within MOMENTUM are as proxies for the purchasing power of people living in a given area when other, more relevant data are not available or to complement existing data.

7.7 Business statistics

7.7.1 Introduction

Starting with **Madrid**, the city council publishes every six months the census of premises and activities. The last update was in July 2019. It provides information about the commercial premises located in each district and neighbourhood with information about the access type (external/internal) and activity. The dataset can be downloaded from the statistics portal of the Madrid City Council[1] in Excel format. The temporal scope of this data source is July 2018 -January 2020.

In **Regensburg** case study, the dataset was obtained again from the "Statistical Yearbook" described in Section 7.1.1. In this case, the source is the Bavarian State Office for Statistics. For this subcategory, the information

[1]https://www.madrid.es/portales/munimadrid/es/Inicio/El-Ayuntamiento/Estadistica/Areas-de-informacion-estadistica/Economia/Empresas-y-locales/Censo-de-Locales-y-Actividades/?vgnextfmt=default&vgnextoid=a3f3fda25c2dc310VgnVCM1000000b205a0aRCRD&vgnextchannel=bfea4f7c93e1a210VgnVCM1000000b205a0aRCRD

available (Chapter 4) is the number of business per industrial sector, total sales, number of employees, etc. The temporal scope is 2014-2018, and the geographical scope and granularity are the City of Regensburg.

Regarding **Leuven**, the dataset available come from the Belgian Crossroads Bank of Enterprises (KBO), and it allows us to obtain information about the number of businesses per activity category for each statistical sector. It only provides data on active registered entities, and the last update was made in November 2018. The dataset can be obtained from the KBO's open data portal[1] in CSV format.

At the moment of writing this deliverable, **Thessaloniki** lacks information about business statistics. However, CERTH will explore the possibility of using business data from ICAP[2] databases for this purpose.

7.7.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Census of Premises and Activities	Regional Statistics Agency	National Crossroads Bank of Enterprises	N/A
Reliability		High	High	High	N/A
Temporal & Geographical Scope	Last update	Jan 2020	December 2019	2019	N/A
	Temp scope	Jul 2018 – Jan 2020	1998 - 2019	2019	N/A
	Geo scope	City	City	Country	N/A
Gran. & Level of Detail	Temp gran	Yearly	Yearly	Yearly	N/A
	Geo gran	Neighbourhoods	City	Exact location	N/A
	Level of Detail	Aggregated measures	Aggregated measures	Individual Business Information	N/A
Availability & Accessibility		Whole project Excel All partners	Whole project PDF, Excel All partners	Whole project csv, excel, gpkg All partners	N/A
Relevance/Usability		High	Low	High	N/A

Table 30. General overview of the Data Quality Assessment of Business statistics

[1] https://economie.fgov.be/en/themes/enterprises/crossroads-bank-enterprises/services-everyone/cbe-open-data
[2] https://www.icap.gr/Default.aspx?lang=2

7.7.3 Identified issues

For **Madrid** and **Leuven** case studies, the datasets described do not present relevant issues.

Regarding the data source for Regensburg, the main issue is similar to that mentioned in previous sections, since the geographical granularity available for these statistics is at the city level.

7.7.4 Relevance/Usability

The relevance of the data sources available for **Madrid** and **Leuven** is high, whereas for Regensburg is low because of its coarse-grained granularity. The potential uses of these data sources within MOMENTUM are the inference of the activities associated with trips, especially in the case of Leuven because of its fine-grained geographical granularity. Another possibility, it is the study of the influence of the business activity in the adoption of the shared-mobility services.

This section aims to comment on other socio-demographic data sources that do not fall into any of the above categories but which may also be relevant to MOMENTUM and to provide information on repositories where other data sources can be found if necessary.

7.8 Other socio-demographic data sources

For **Madrid**, **Regensburg** and **Thessaloniki**, no relevant data sources have been identified other than those mentioned above. In case that needs are identified that are not covered by the data sources set out here, the possible repositories of information to which to turn in the case of Madrid would be the statistics section of the Madrid City Council website or the open data portal of the National Statistics Institute[1]. In the case of Regensburg, it would be the "Statistical Year Book". And in the case of Thessaloniki, it would be the city council's open data portal [2].

For **Leuven**, another source of relevant socio-demographic data does exist, which is the Student Housing Inventory. As explained in Section 7.1, students are by law not registered in Leuven but their parents' city. As in Leuven the proportion of students is really high (around one-third of the real population) without estimates of the number of students in each area, the picture of the city would be partial. From the above-mentioned data source, Leuven has obtained the estimation of the number of unregistered students in each statistical sector. The repositories for other sources of socio-demographic data would be the open data portal of Statbel[3], KBO[4] and KSZ[5].

[1] https://www.ine.es/ss/Satellite?L=0&c=Page&cid=1259942408928&p=1259942408928&pagename=ProductosYServicios%2FPYSLayout
[2] https://opendata.thessaloniki.gr/el
[3] https://statbel.fgov.be/en/open-data
[4] https://economie.fgov.be/en/themes/enterprises/crossroads-bank-enterprises/services-everyone/cbe-open-data
[5] https://www.ksz-bcss.fgov.be/nl/dwh/homepage/index.html

8 Data Quality Assessment for Travel Time Data Sources Travel Time Data

8.1.1 Introduction

In the **Madrid Case Study**, the travel time data available come from Google Maps API[1] and have been collected by Nommon. It consists of travel time between different centroids in Madrid at specifics time of the day and days of the week (e.g. Wednesdays at 7:00 AM), for different transport modes. Concretely, the information provided by the dataset includes squared travel time matrices for 412 centroids covering central districts, with a separation of 500 metres between each other. There are five matrices: driving on Wednesday at 7:30 AM; Transit on Wednesday at 7:30 AM; driving on Wednesday at 5:30 PM; Transit on Wednesday at 5:30 PM; and Transit on Wednesday at 12:00 AM.

For the city of **Leuven**, the data source was obtained by the City Council from TomTom traffic data and traffic stats services[2]. It is, therefore, a private dataset that it will be available to all partners in MOMENTUM only for project purposes. It consists on travel time information for 98 different routes in Leuven and the metropolitan area taken on December 6th and 7th 2019 at different times: 6:00, 7:00, 7:30, 7:45, 8:00, 8:15, 8:30, 9:00, 10:00, 11:00, 12:00, 13:00, 14:00, 15:00, 16:00, 16:30, 17:00, 17:15, 17:30, 17:45, 18:00, 18:30, 19:00, 20:00 and 21:00. There are two files in CSV format. The first one refers to route description, and for each route provides a name, the origin and destination GPS coordinates, and a heading angle indicating the orientation. The second file is related to the travel time information and for each route, day and time it contains the travel time in free flow, historically averaged and current conditions.

In the **Thessaloniki Case Study**, the dataset available is obtained from the data captured by iTravel sensors installed in key locations (major road junctions) of Thessaloniki. Those sensors detect Bluetooth-enabled devices (e.g. smartphones, etc.) in their range and record their unique MAC address, provided that the Bluetooth connectivity is enabled and in discoverable mode. The calculation of travel times leverages consecutive detections of the same MAC address at different iTravel sensors. The algorithm is run every 15 minutes and utilizes the sensors' data collected within the previous temporal interval. The iTravel devices are owned by the Hellenic Institute of Transport, Center for Research and Technology Hellas (CERTH-HIT). Each record of this dataset includes a certain path's id, the corresponding timestamp and the temporal duration (travel time) estimated for this path.

No similar datasets are available for **Regensburg**. In case it is necessary, the use of Google Maps API, some tools that make use of Open Street Maps (e.g. OpenTripPlanner[3]) or similar services/tools will be considered.

[1] https://cloud.google.com/maps-platform/routes/
[2] https://www.tomtom.com/products/historical-traffic-stats/
[3] https://www.opentripplanner.org/

8.1.2 General overview of the Data Quality Assessment

Type & Data quality dim.		Madrid	Regensburg	Leuven	Thessaloniki
Type		Google Maps API	N/A	TomTom	Bluetooth devices
Reliability		High	N/A	High	High
Temporal & Geographical Scope	Last update	Continuous updates	N/A	December 7th, 2019	Continuous update
	Temp scope	N/A	N/A	December 6th and 7th, 2019	From 2017
	Geo scope	City (412 Centroids separated by 500m)	N/A	City + Metropolitan area	City
Gran. & Level of Detail	Temp gran	Wednesday at 7:30AM, 5:30AM and 12:AM	N/A	Specific times during the day	Secondly
	Geo gran	Centroid pairs	N/A	Specific predefined paths	Specific predefined paths
	Level of Detail	Average travel time for driving and transit	N/A	Driving travel time	Driving travel time
Availability & Accessibility		Whole project JSON All partners	N/A	Whole project JSON All partners	Whole project JSON All partners
Relevance/Usability		High	N/A	Medium	High

Table 31. General overview of the Data Quality Assessment of Travel time data

8.1.3 Identified issues

Regarding the dataset available for **Madrid**, the main issues identified are, on the one hand, only data on Wednesdays are available, due to the cost of Google Maps API, which could lead to biases towards the special

characteristics of this weekday in Madrid; and on the other hand, the only availability of travel time for public transport and driving, and not for others as pedestrian or bike.

In the case of the data source available for **Thessaloniki**, the only issue is similar to the previous one as only travel times for driving are measured. We would like to mention that the Bluetooth detectors capture all devices, even the ones used by pedestrians. In this sense, pedestrian travel times could be captured, but this is something that has not been accomplished so far.

As for **Leuven**'s dataset, on the one hand, it presents the same issue as Thessaloniki, and on the other hand, only travel time information from two days is available which poses a high risk of bias when that information is considered. However, having information on historical averages of travel times reduces the impact of this risk.

8.1.4 Relevance/Usability

The relevance of the two data sources available in **Madrid** and **Thessaloniki** is high, mainly because of their good geographical scope, granularity and timeliness, whereas for **Leuven** is medium because of the short temporal scope. The potential uses of these data sources within MOMENTUM are to study the relation between the accessibility of some zones by some transport modes (measured in terms of travel time from/to specific locations) and use of shared mobility services or public transport.

9 Conclusions

In this document, we have provided an inventory **with a total of more than 80 data sources** available for MOMENTUM. They were classified into five main categories: transport supply, transport demand, maps & cartography, socio-demographic and travel time. For each of these five main categories, different sub-categories were also identified. Apart from this, a **data quality assessment** was also performed to each identified data source in terms of reliability, sample size, geographical and temporal scope, geographical and temporal granularity, completeness, validity and accessibility. From the results of the previous data quality assessment, we also accomplished an **analysis of the potential usability** that each of these data sources may have for MOMENTUM according to its characteristics.

From the previous analyses, we concluded that the **most relevant data sources for each case study**, and therefore, with higher usability in the next tasks of WP3 and upcoming WPs, are the following:

Madrid Case Study

- **Transport supply**
  - Schedules and lines for Underground, Urban buses, Interurban buses and Suburban trains
  - Bike-sharing supply data from BiciMAD (station-based)
  - Public off-street parking supply data form EMT
- **Transport demand**
  - Smart Card Validations for urban buses
  - Bike-sharing demand data from BiciMAD (station-based)
  - Cycling count from fibre-optics sensors
  - Pedestrian counts from fibre-optics sensors
  - Traffic data from loop sensors
  - Public off-street parking demand data
- **Maps & Cartography**
  - Land use GIS database
  - Weather data
- **Socio-demographic**
  - Demographic data from the National Statistics Agency
  - Income statistics from the National Tax Income Database
  - Car ownership from the Regional Vehicle Tax Database
  - Business statistics from regional database
- **Travel times**
  - Travel times for driving and transit from Google Maps API

Regensburg Case Study

- **Transport supply**
  - Schedules and lines for urban buses
  - Time deviations from planned schedules in urban buses
  - Station-based car-sharing supply data
- **Transport demand**
  - Passenger counts from sensorised buses
  - Household survey from 2018
  - Station-based car-sharing demand data
- **Maps & Cartography**
  - Land use information from the GIS database

  - Weather data from DarkSky.net or Weatherunderground
- **Socio-demographic**
  - Demographic data from the "Statistics Yearbook"
  - Labour market data from the "Statistics Yearbook"

Leuven Case Study

- **Transport supply**
  - Schedules and lines for urban buses
  - Road network database from Wegen Register
  - Bike-sharing supply data from Blue Bikes service (station-based)
  - Public off-street and shop&go parking supply data
- **Transport demand**
  - Bike-sharing demand data from Blue Bikes service (station-based)
  - Cycling counts from fibre-optic sensors
  - Pedestrian counts from Wifi detectors
  - Cycling, pedestrian and traffic counts from Telraam
  - City Monitor Survey
  - Students Survey
  - WWV Survey for commuters
  - Public off-street parking demand data
- **Maps & Cartography**
  - Land use data from City GIS database
  - Weather data from DarkSky.net or Weatherunderground
- **Socio-demographic**
  - Demographic data from the National Statistics Agency
  - Income data from Tax Income Database
  - Car-ownership data from City Monitor Survey
  - Labour market statistics from Crossroads Bank for Social Security
  - Business statistics from Belgian Crossroads Bank of Enterprises
- **Travel time**
  - Driving travel times from TomTom API

Thessaloniki Case Study

- **Transport supply**
  - Schedules and lines for buses
  - Transport Network Database
  - Taxi trip data from Taxiway
  - Bike-sharing service supply data from Thessbike (station-based)
  - E-scooter-sharing service supply data from Lime and Hive (free-floating)
- **Transport demand**
  - Bike-sharing service demand data from Thessbike (station-based)
  - Household survey from 2017-2018
  - E-scooter-sharing service demand data from Lime and Hive (free-floating)
  - Floating-car traffic data from Taxiway
  - Taxi trip data from Taxiway
  - Geo-tagged check-in events from Facebook
- **Maps & Cartography**
  - Land use City GIS database

    - Weather sensors
- **Socio-demographic**
    - Demographic data from 2011 population census
    - Income statistics from 2017-2018 household survey
    - Car ownership from 2011 population census
    - Labour market statistics from 2011 population census
- **Travel-time**
    - Driving travel times from iTravel Bluetooth devices

10 Appendix A. Completeness, Validity and Simple Exploratory Analysis of relevant data sources

10.1 Description of the completeness and validity measures considered

Before describing the results of the analyses performed for the different data sources considered in this appendix, we will describe the two completeness and validity measures considered for this analysis, that are %MIV (percentage of Missing or Invalid Values) and CPwMD (Continuous Periods with Missing Data):

- **%MIV:** it corresponds to the total number of missing or invalid values in the dataset. The formula would be the next one: $\%MIV = \frac{SumMIV}{NS \times number\ of\ fields\ in\ the\ sample\ record} \times 100$, where NS is number of samples in the dataset and $SumMIV = \sum_i^{NS} numMIV(sample(i))$, where $sample(i)$ refers to the data entry of the i-th sample, and $numMIV(sample)$ corresponds to the number of missing or invalid values of the sample.
- **CPwMD:** it refers to the duration of the periods in which all samples present a high percentage of missing data, in such a way that the information contained during that period cannot be used. We will also refer to those periods as periods with lost data. It can be due to a high percentage of missing data or to the absence of data.

10.2 Madrid Case Study

10.2.1 Public Transport Smart Card Data

This section refers to the study of the PT demand data source described in Section 5.1 for Madrid. In this case, we performed completeness, validity and exploratory analysis with the sample data provided that corresponds to seven days, concretely, from November 15th, 2019 till November 21st, 2019.

10.2.1.1 Completeness and Validity Analysis

The %MIV and CPwMD analyses concluded that there are no missing data nor periods without data in this case.

10.2.1.2 Exploratory Data Analysis

The plot below shows the number of users per hour of the day. Along the period with sample data, more than one million made a trip in EMT buses. We can observe that there are three peaks, one at 8 am, one at 2 pm and another one between 5 pm and 6 pm, which can be associated with the commuting activity.

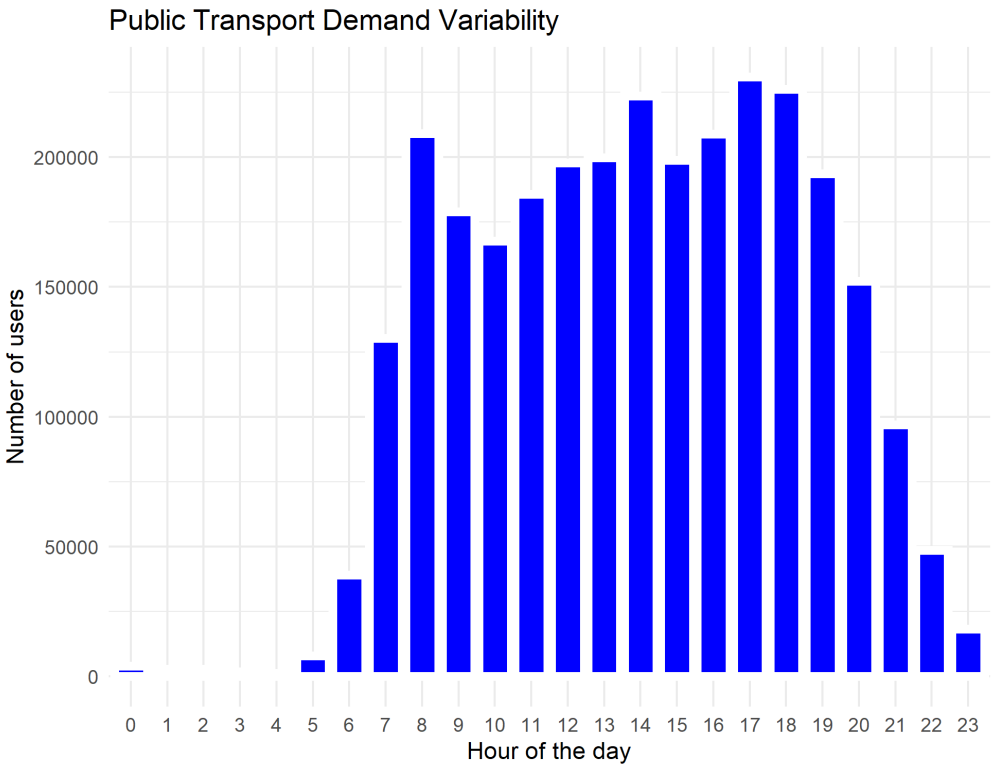


Figure 2 Madrid Case Study: Public Transport Demand Temporal Variability

10.2.2 Bike-Sharing Demand

This section refers to the analysis of the data source described in Section 5.3 for Madrid, concretely the bike-sharing data provided by BiciMAD. In this case, we performed completeness, validity and exploratory analysis with data from April 1st, 2019 till June 26th, 2019.

10.2.2.1 Completeness and Validity Analysis

The analysis of this dataset concluded that the %MIV is 10% for half of the samples. A more in-depth analysis shows that the missing data in all of these cases is the field with the zip code of the user, which means the zip code of the user's residence is available for half of the trips. Regarding the CPwMD, the figure below shows the distribution of periods without data with a duration lower than 50 hours per bike-sharing unplugged station. We can see that the majority of the periods without data have a duration of 5 hours or less with a peak in one hour. There are only 25 periods without data with a duration higher than 50 hours, being the longest period of 265 hours. In this way, we can conclude that the completeness and validity of this data source is not an issue.

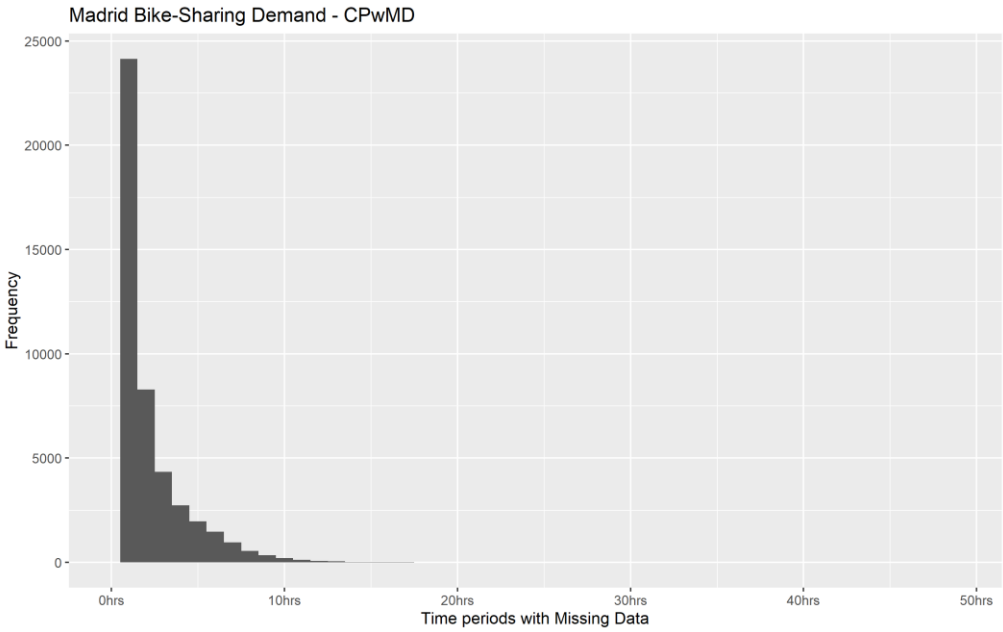


Figure 3 Madrid Case Study: Bike-Sharing Demand - CPwMD

10.2.2.2 Exploratory Data Analysis

The bike-sharing demand variability per hour of the day and day of the week shows that there are two peak-times for the use of this service during weekdays, one in the morning from 8 to 9 am, and one in the evening from 6 to 7 pm. Interestingly, on Fridays the evening peak time move to 3 pm, a usual time to leave the workplace on that day.

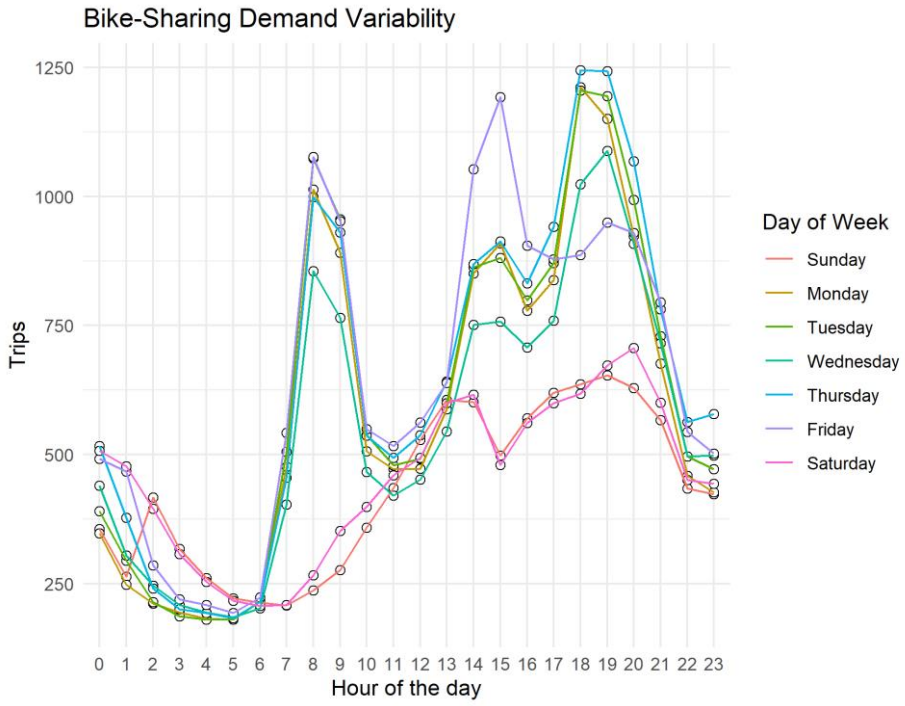


Figure 4 Madrid Case Study: Bike-Sharing Demand Temporal Variability

The chart below shows the trip duration distribution for BiciMAD bike-sharing service. We can see that the mode of the distribution is located around 6-7 minutes and that most of the trips are shorter than 20 minutes (80 per cent).

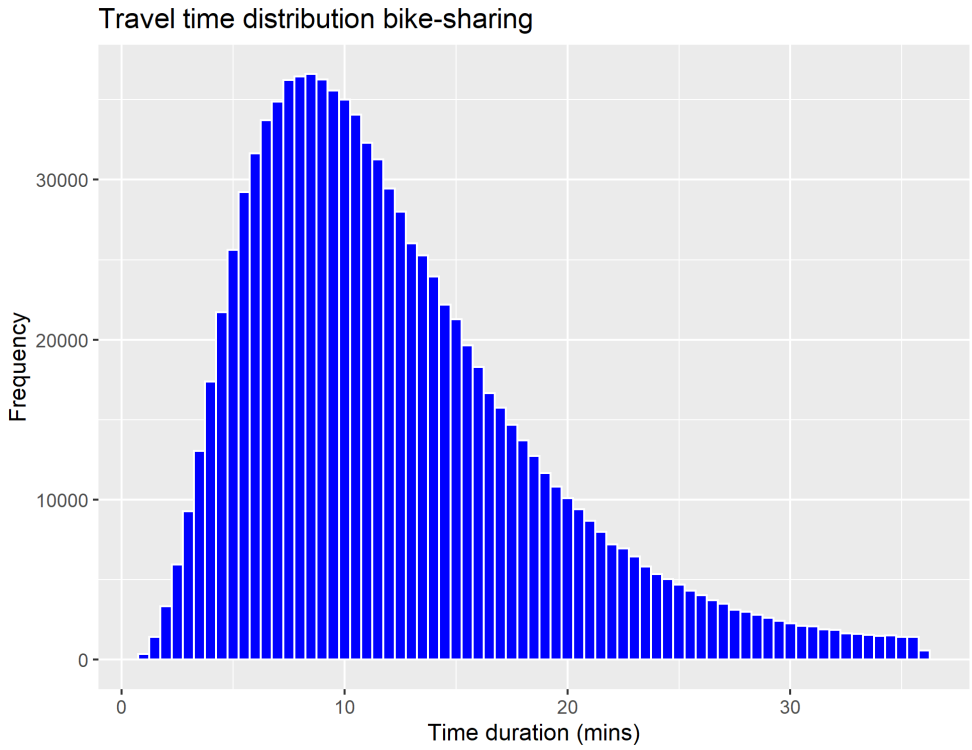


Figure 5 Madrid Case Study: Bike-sharing Trip Duration Distribution

10.2.3 Cycling Demand

This section refers to the analysis of the cycling demand data source described in Section 5.4 for Madrid. In this case, we performed completeness, validity and exploratory analysis with data from January 1st, 2019 till June 30th, 2019.

10.2.3.1 Completeness and Validity Analysis

The analysis of the %MIV and CPwMD for this dataset concluded that there are no missing data and no periods without data.

10.2.3.2 Exploratory Data Analysis

The cycling demand variability per hour of the day and day of the week shows large differences among weekdays and more frequent use of bikes during weekends and Fridays, suggesting that in Madrid the use of the bike may be more associated to leisure or sportive activities.

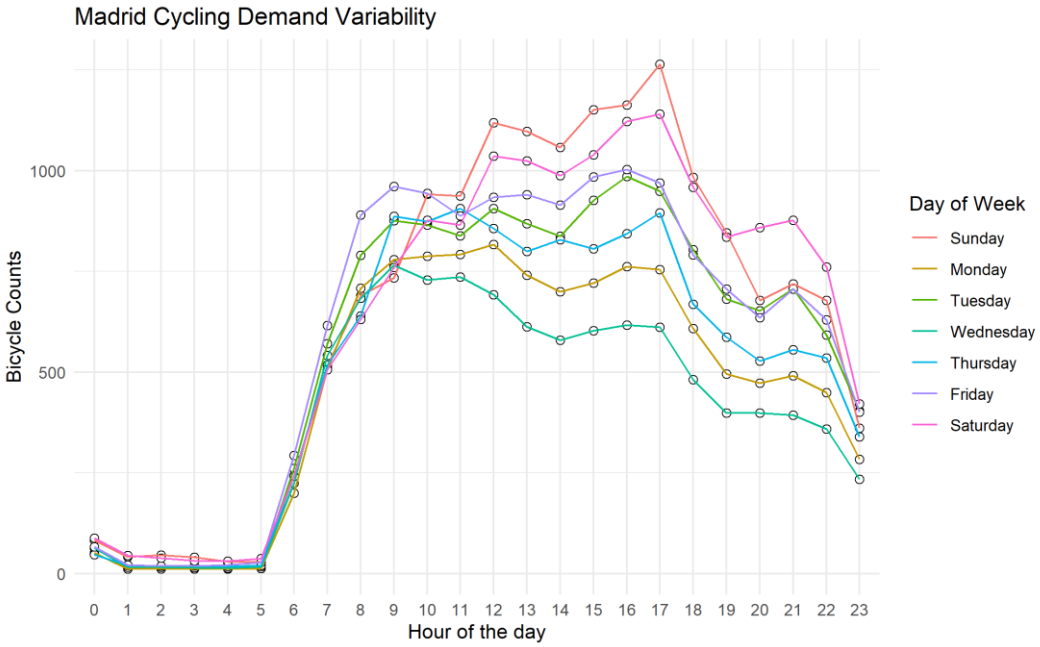


Figure 6 Madrid Case Study: Cycling Demand Temporal Variability

10.2.4 Pedestrian Demand

This section analyses the pedestrian demand data source for Madrid described in Section 5.5. In this case, we performed completeness, validity and exploratory analysis with data from January 1st, 2019 till June 30th, 2019, as in the previous section.

10.2.4.1 Completeness and Validity Analysis

The analysis of the %MIV and CPwMD for this dataset concluded that there are no missing data and no periods without data.

10.2.4.2 Exploratory Data Analysis

The figure below displays pedestrian demand variability per hour of the day and day of the week. We can check that walking is more frequent during weekends and Fridays, which can also be linked to tourism. In this case, there are no clear peak-times but a period that concentrates most of the pedestrians between 12 pm and 7 pm.

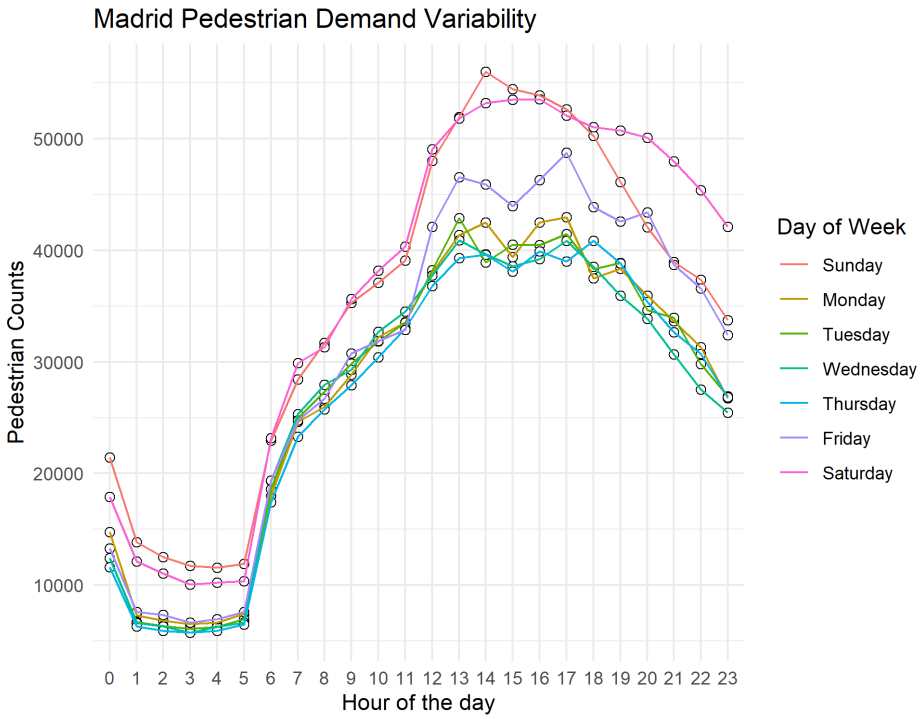


Figure 7 Madrid Case Study: Pedestrian Demand Temporal Variability

10.2.5 Telecom Data

This section refers to the analysis of the data source described in Section 5.7 on Call Detail Records with data collected on October 15th, 2019. In this case, we only performed an exploratory data analysis.

10.2.5.1 Exploratory Data Analysis

The number of users connected to Orange Network that day was more than 18 million and generated more than 1500 million phone records. The graph below shows the distribution of the percentage of users with some phone activity and the percentage of phone records during the reference day. We can observe that during the night the rate of users with phone activity drops to 30%, but during the day it keeps higher than the 60%. The distribution of records follows a similar pattern.

In the next plot, we show different percentiles for the number of records generated on average for each user. It can be seen that during the daytime, most of the users produce eight records per hour, but this number drops to 2 during the night.

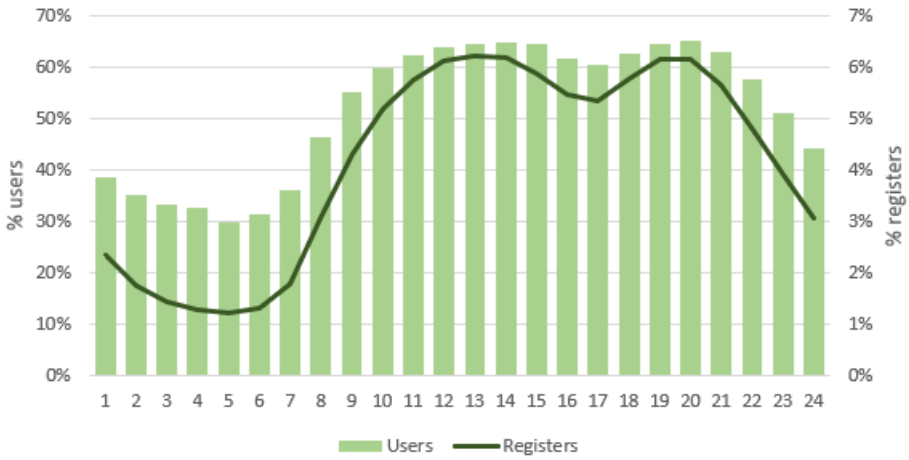


Figure 8 Hourly distribution of mobile phone users and records

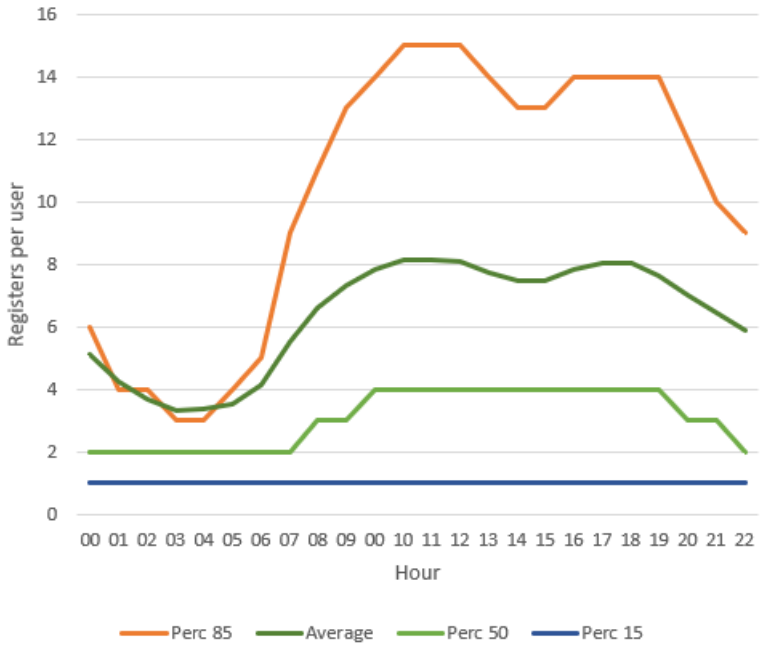


Figure 9 Distribution of records per user in each time interval

In the next plot, we show the probability density function of time intervals between consecutive data connections of the user's mobile phones. Interestingly, we can observe peaks at 30 and 60 minutes that are usually generated by mobile apps that perform automatic updates at these intervals.

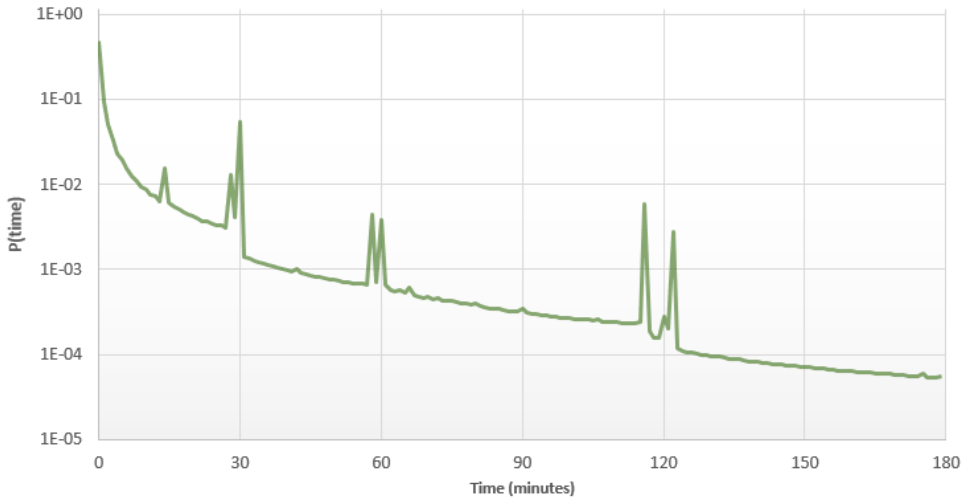


Figure 10 Probability density function of time intervals between consecutive data sessions

Another interesting analysis regarding this data source is its spatial accuracy, particularly for the city of Madrid, as it depends on the spacing between towers providing coverage in the area. In this case, the radius of coverage of more than 50% of the towers is less than 300 meters (see Figure 11) which allows characterizing medium and long-distance trips. Furthermore, this distribution is spatially heterogeneous with a high density in the city centre whereas in the outer parts the radiuses are larger (see Figure 12).

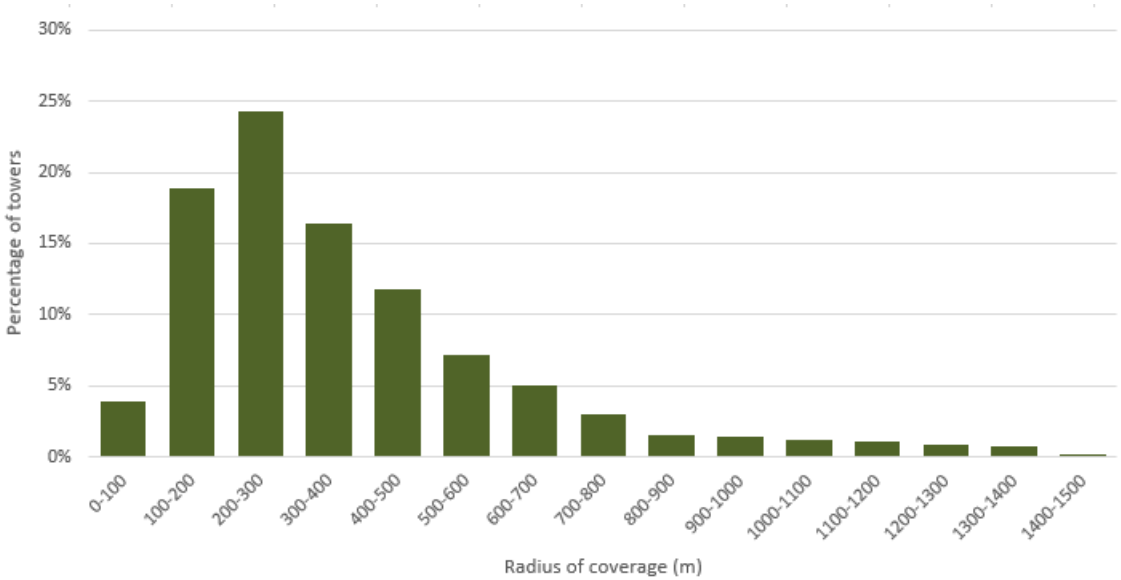


Figure 11 Distribution of the radius of coverage for Madrid cell towers

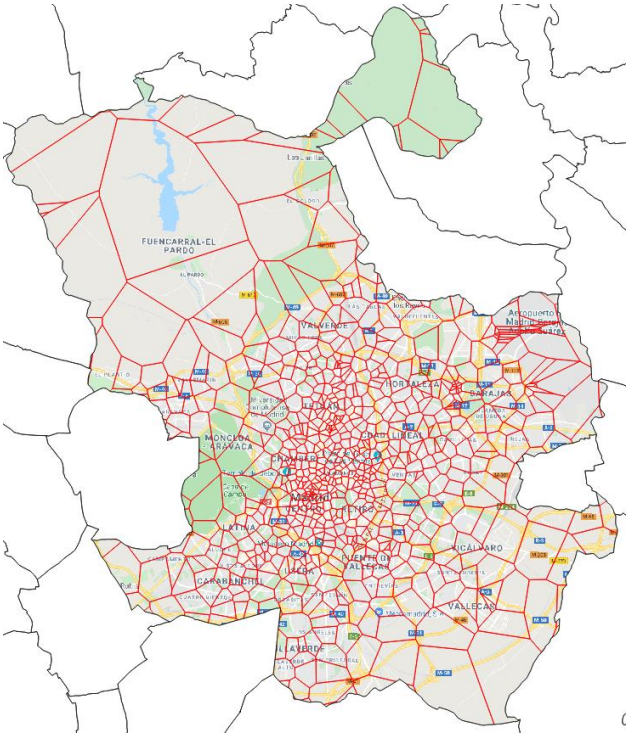


Figure 12 Coverage areas for the City of Madrid

To finish with this data source, we have compared the population pyramid of Orange's users and for Spain, according to the Spanish National Statistics Agency (INE, by its Spanish acronym) displayed in Figure 13. In this case, we can see the limitations of the representativeness of this data source. People between 0 and 24 is almost inexistent (either because they do not own a mobile phone or because it is registered on their parents 'name). In contrast, the middle ages (between 35 and 65) are overrepresented. This is an important consideration when expanding the sample.

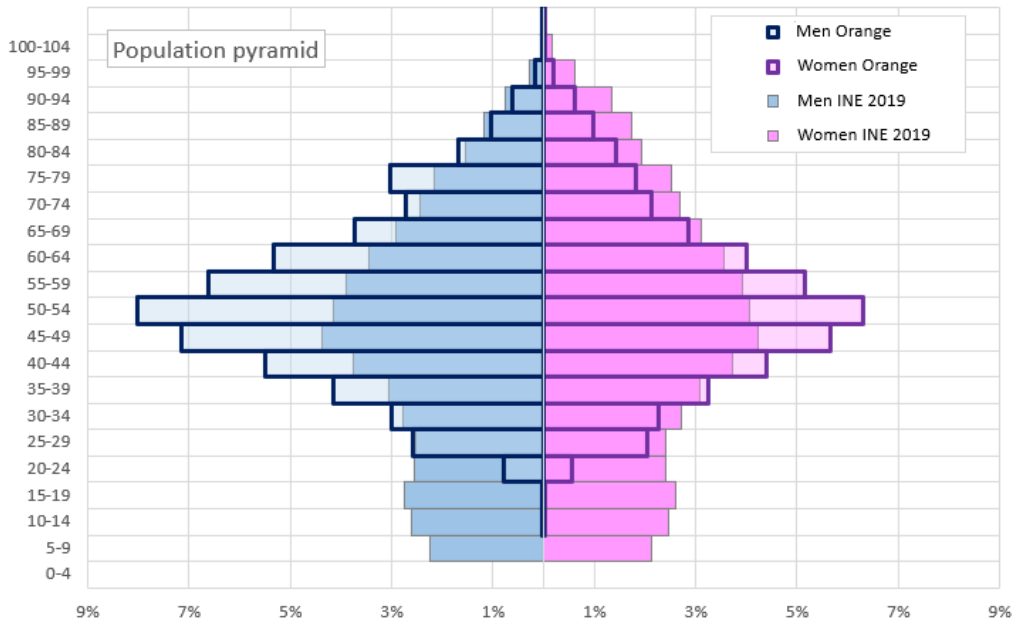


Figure 13 Population pyramid for Spain and Orange Spain users

10.2.6 Traffic Data

This section refers to the analysis of the traffic data source for Madrid described in Section 5.9. In this case, we performed completeness, validity and exploratory analysis with data from October 1st, 2019 till December 31th, 2019.

10.2.6.1 Completeness and Validity Analysis

The analysis of this dataset concluded that the %MIV is 0, that is, there are no missing data on those time slots with at least some available data. Regarding the CPwMD, the figures below show the distribution of periods without data, where the periods are shorter than 100 hours and longer, respectively. Focusing on the first plot, we can see that in most of the cases the duration is very short (one or two hours), so it does not present a relevant problem for the analysis. Considering periods longer than 100 hours, it can be observed that most of them show a duration that ranges between 100 and 150 hours, that is, between 4 to 6 days approx. It is important to note that the frequency is low, as in most cases, only one or two periods are found. Some relevant periods reach more than 400 hours (15 days), but they are very infrequent. Having said that, we can conclude that CPwMD, in general, does not pose an issue for this data source, although it will require some data cleaning to remove those sensors with long periods without data.

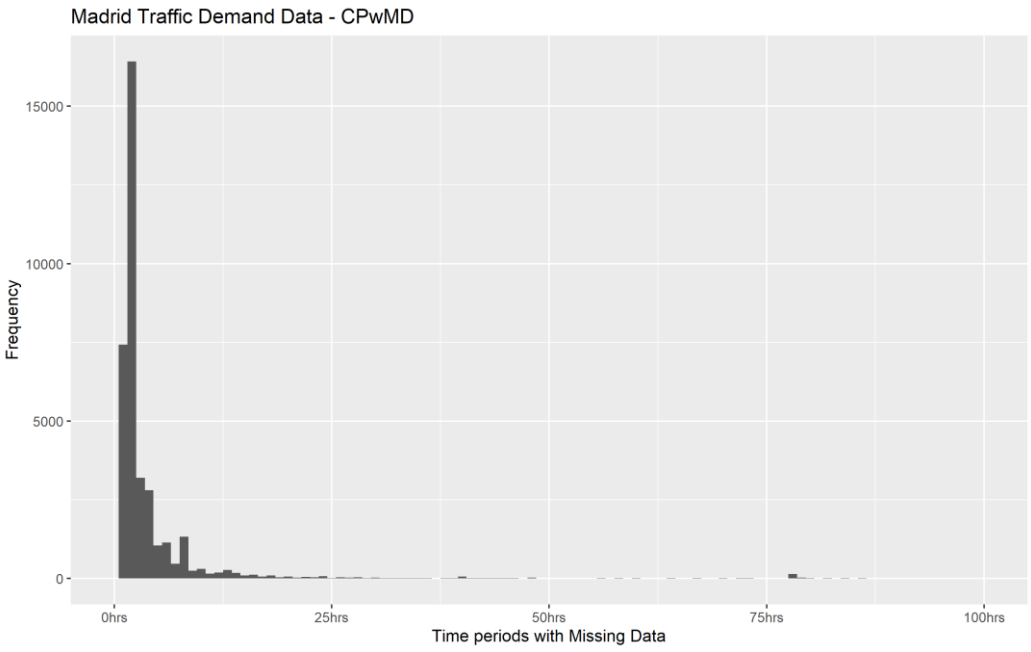


Figure 14 Madrid Case Study: Traffic data -CPwMD I

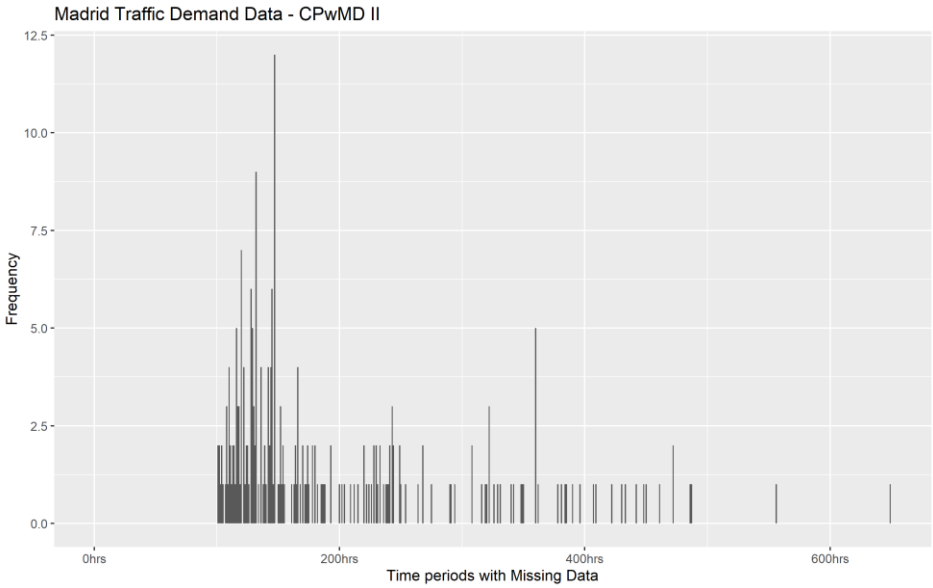


Figure 15 Madrid Case Study: Traffic Data -CPwMD II

10.2.6.2 Exploratory Data Analysis

The traffic demand variability per hour of the day and day of the week show patterns related with commuting with a morning peak from 7 to 9 am, and an evening peak at 6 pm, as it can be observed below. Traffic during weekends is also noticeably less.

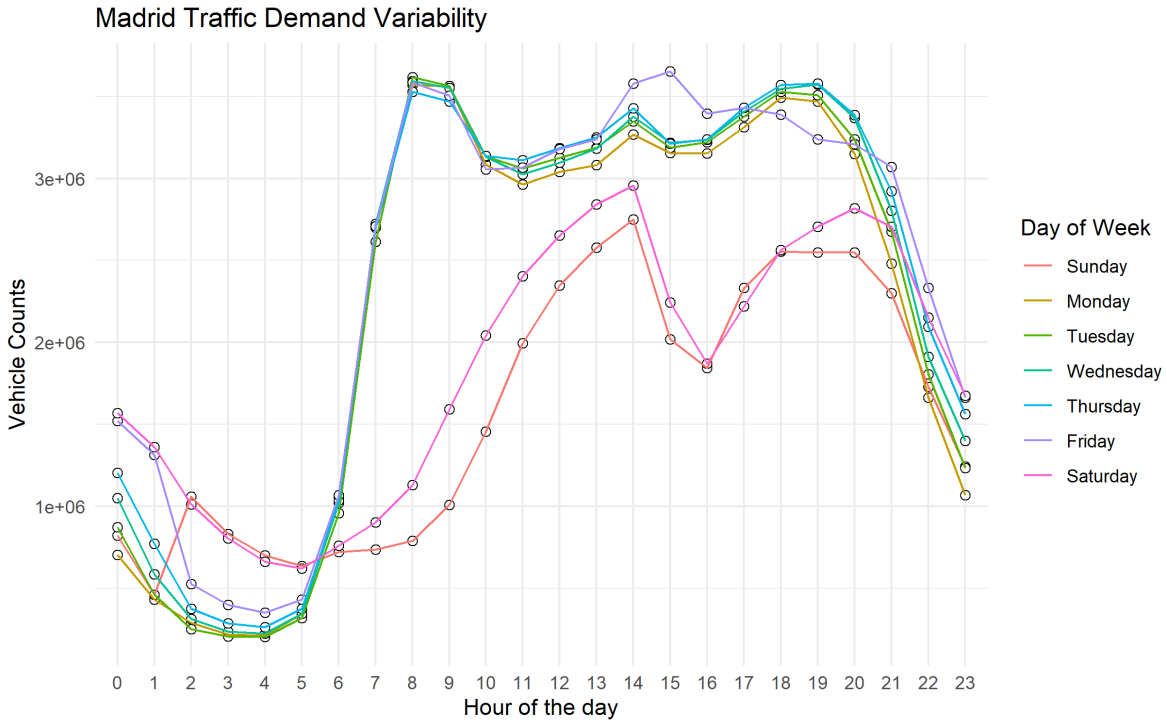


Figure 16 Madrid Case Study: Parking Demand Temporal Variability

10.2.7 Parking Demand

This section refers to the analysis of the parking demand data source described in Section 5.12 for Madrid. In this case, we performed completeness, validity and exploratory analysis with data from the three available parking places with historical data from September 1st, 2019 until October 31th, 2019.

10.2.7.1 Completeness and Validity Analysis

The analysis of this dataset concluded that the %MIV is 0, that is, there are no missing data on those time slots with at least some available data. Regarding the CPwMD, the figure below shows the distribution of periods without data per parking. We can see that, from the point of view of the completeness, this data source does not present any issue for data analysis. The most prolonged periods are no longer than 30 hours and the corresponds to Saturday nights and Sundays, where some parking facilities are not opened therefore, they do not report any data.

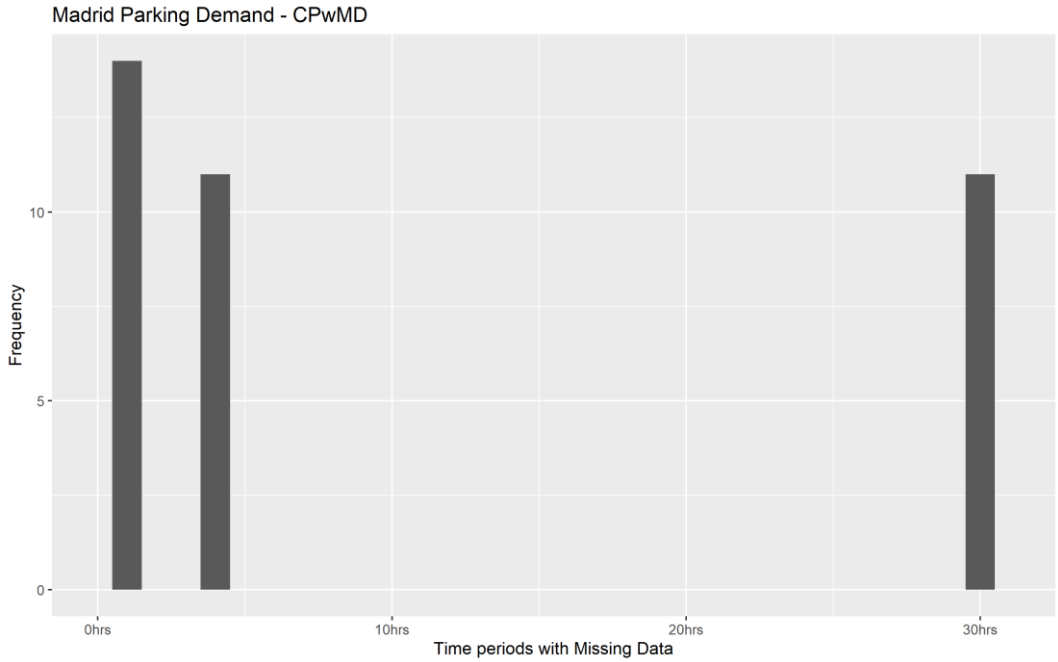


Figure 17 Madrid Case Study: Bike-Sharing Demand -CPwMD

10.2.7.2 Exploratory Data Analysis

The parking demand variability per hour of the day and day of the week shows that in weekdays the occupancy of parking reaches its peak between 11 am and 1 pm, whereas in weekends there two peaks, one at 3 pm and another one on Saturdays at 10 pm.

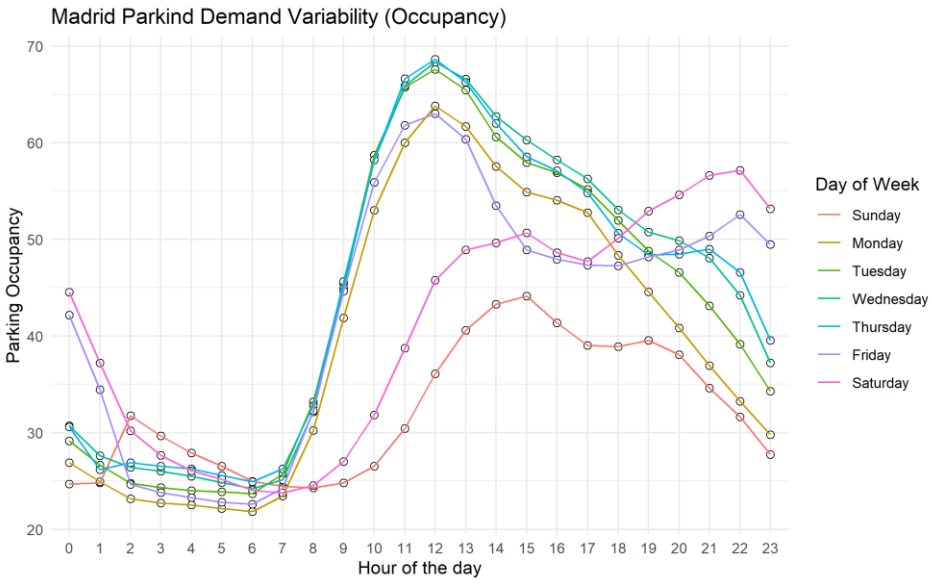


Figure 18 Madrid Case Study: Parking Demand Temporal Variability

10.3 Regensburg Case Study

10.3.1 Cycling Demand

This section refers to the analysis of the data source described in Section 5.4 for cycling demand in Regensburg. In this case, we only performed an exploratory analysis as the data comes from a frequency survey.

10.3.1.1 Exploratory Data Analysis

In this case, we have analysed the variability of cycling demand per hour of the day and day of the week. The plot below shows how the use of the bike in Regensburg is higher during the afternoon in weekdays, and during the morning on Saturdays. This suggests that it is mostly used for sportive or leisure activities.

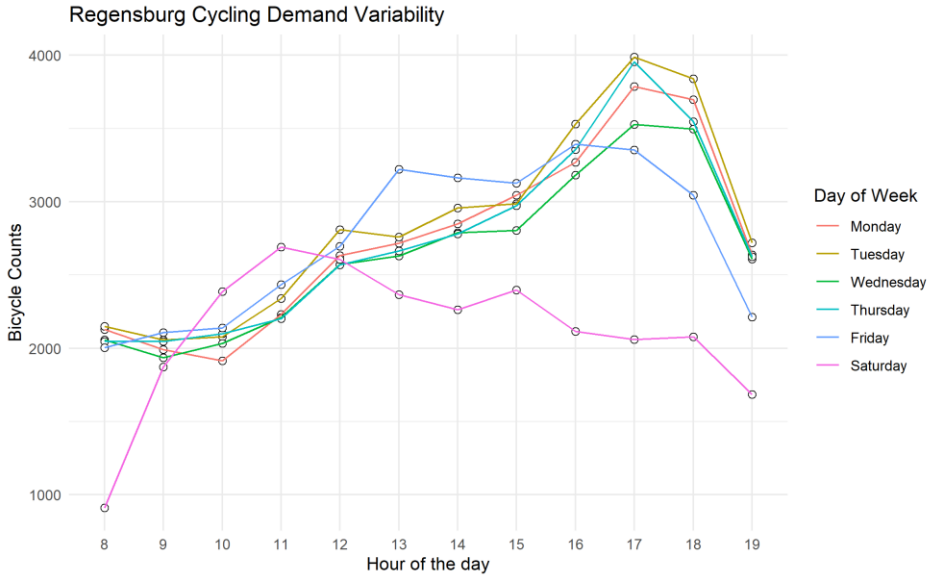


Figure 19 Regensburg Case Study: Cycling Demand Temporal Variability

10.3.2 Pedestrian Demand

This section refers to the study of the data source described in Section 5.5 for pedestrian demand in Regensburg. In this case, we only performed an exploratory analysis as the data comes from a frequency survey.

10.3.2.1 Exploratory Data Analysis

We have analysed the variability of pedestrian demand per hour of the day and day of the week. The plot below shows how similar trends for weekdays with two peaks at 12 pm and 5 pm (4 pm for Fridays), respectively, whereas on Saturday, the number of pedestrians is much higher, reaching the peak between 2 pm and 4 pm. This is mostly due to tourism activity as the frequency survey took place in Regensburg's old town.

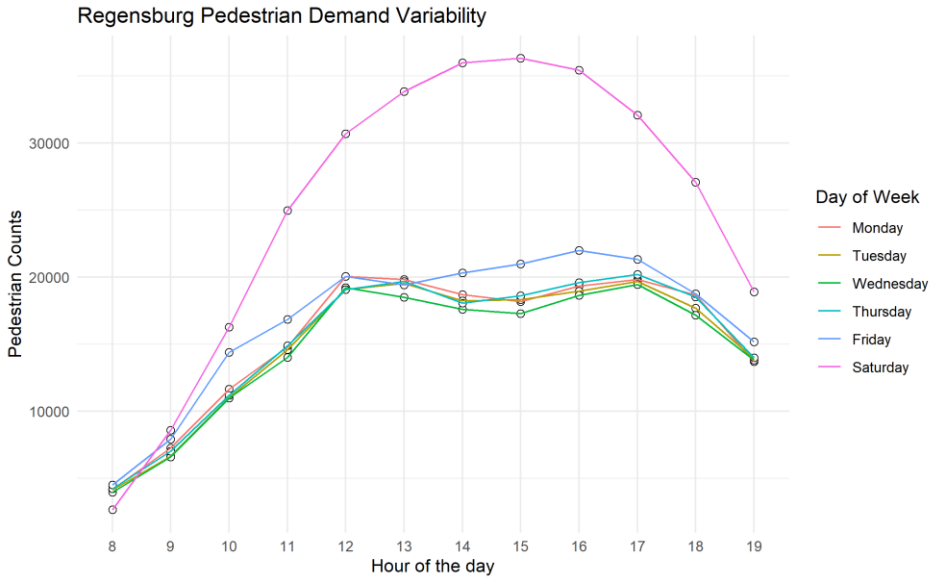


Figure 20 Regensburg Case Study: Pedestrian Demand Temporal Variability

10.3.3 Car-sharing Demand

This section refers to the study of the data source described in Section 5.8. In this case, we performed completeness, validity and exploratory analysis, using all historical data available, that is, from November 15th, 2016 till November 28th, 2019.

10.3.3.1 Completeness and Validity Analysis

The completeness and validity analysis of this data source establishes that the %MIV for all sensors is approximately 5% in only 5% of the samples. These missing values are concentrated in the field that provided the trip length, which is a relevant field. Still, taking into account the low percentage of samples, it does not present an issue for the usability of this data source.

10.3.3.2 Exploratory Data Analysis

In this case, we have analysed the variability of cycling demand per hour of the day and day of the week. The plot below shows that in general, the use of this service is low since the average does not reach the two trips per hour at any time of the day. The peak periods for this service are not very clear, but we can see one in the morning between 7 am, and 9 am, and another one in the afternoon between 4 pm and 5 pm, especially on weekdays.

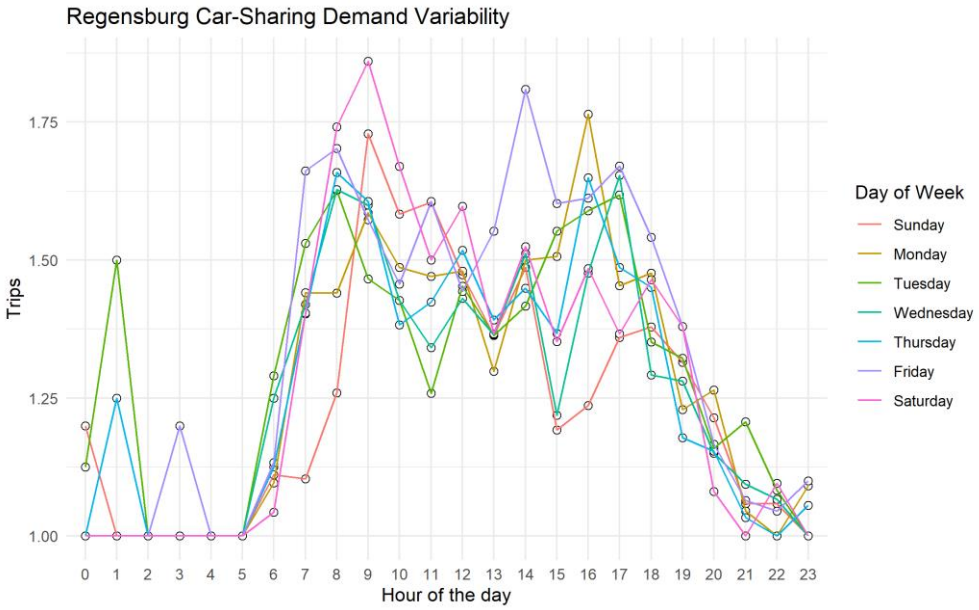


Figure 21 Regensburg Case Study: Car-Sharing Demand Temporal Variability

The chart below shows the trip length distribution (in km) for the car-sharing service available in Regensburg. It is important to note that this considers the length of the whole tour performed by the user since he/she pick-up the car until he/she returns it. In this way, the trip length should be around half of the total kilometres showed in the plot. We can see that a high number of these trips are short (less than 25 km, or 12 km considering only one way), although there is an important portion of the trips in the range between 25kms and 100kms, suggesting that it is also used for medium distance trips.

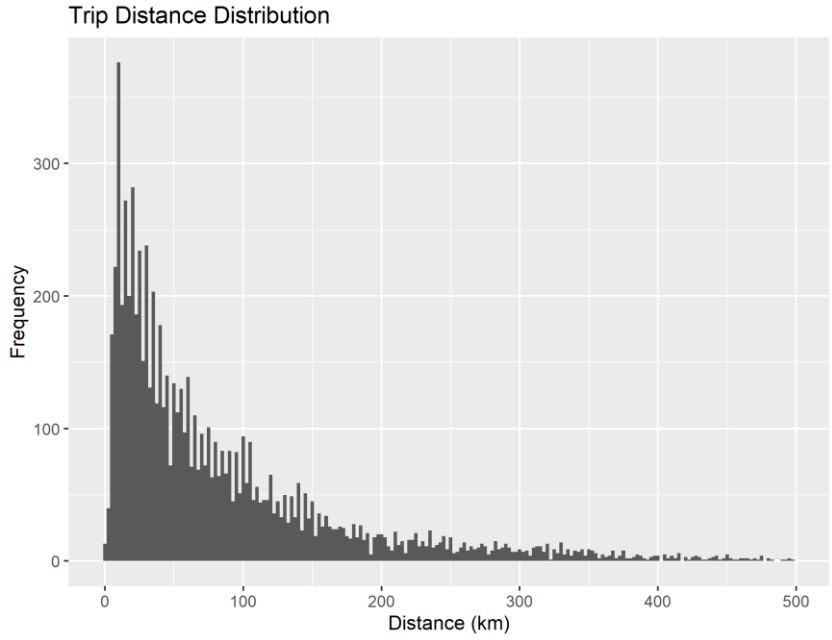


Figure 22 Regensburg Case Study: Car-sharing Trip Distance Distribution

10.4 Leuven Case Study

10.4.1 Bike-sharing Demand

This section refers to the analysis of the data source described in Section 5.3 for Leuven. In this case, only an exploratory analysis of the data has been carried out since, due to the simplicity of the dataset, it is easy to check that there are no missing data.

10.4.1.1 Exploratory Data Analysis

The following graph shows the number of rides registered with the Blue-bikes bike-sharing service in each week of the year 2017. It can be observed how the number of trips has an increasing trend at the end of the year and that it decreases in the holiday periods, mainly Christmas (weeks 14-15), Easter (week 22) and Summer (weeks 28-33).

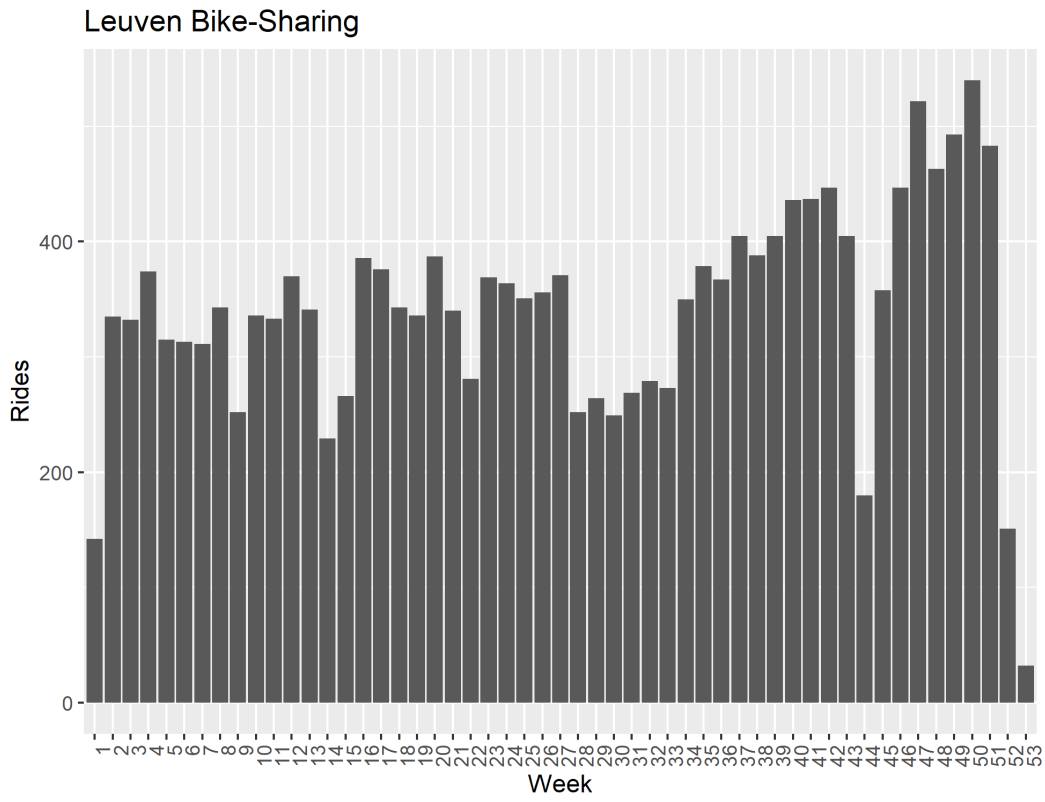


Figure 23 Leuven Case Study: Bike-sharing weekly variability

10.4.2 Cycling Demand

This section refers to the analysis of the data source described in Section 5.4 for cycling demand in Leuven. In this case, we performed completeness, validity and exploratory analysis.

10.4.2.1 Completeness and Validity Analysis

The completeness and validity analysis of this data source establishes that the %MIV for all sensors is 0%, so no missing data have been identified. Regarding the CPWMD, in this case, it is relevant since it can reach up to five

months, probably because of problems in the sensors. In this case, we do not show the distribution as the longest CPwMD are the only relevant issue of the data source. Concretely, the longest CPwMD for each of the five sensors is the following:

- Parijsstraat fietstelweek: from 2015-09-01 05:00:00 to 2015-09-01 12:00:00
- Martel: from 2017-01-16 05:00:00 to 2017-01-21 04:00:00
- Kardinal Mercier: from 2017-05-11 05:00:00 to 2017-11-07 04:00:00
- Brugbergpad: from 2017-04-14 05:00:00 to 2017-10-08 04:00:00
- Aarschotsesteenweg: from 2017-01-17 05:00:00 to 2017-07-01 04:00:00

We can find other continuous periods without data, but they do not present an issue because they are lower than ten hours, and their frequency is also low (up to 5 times per sensor).

10.4.2.2 Exploratory Data Analysis

In this case, we have analysed the variability of cycling demand per hour of the day and day of the week. The plot below shows how the use is significantly higher on labour days than on weekends and that there are two peak hours, one in the morning around 8:00 am, and another in the afternoon, around 5:00 pm.

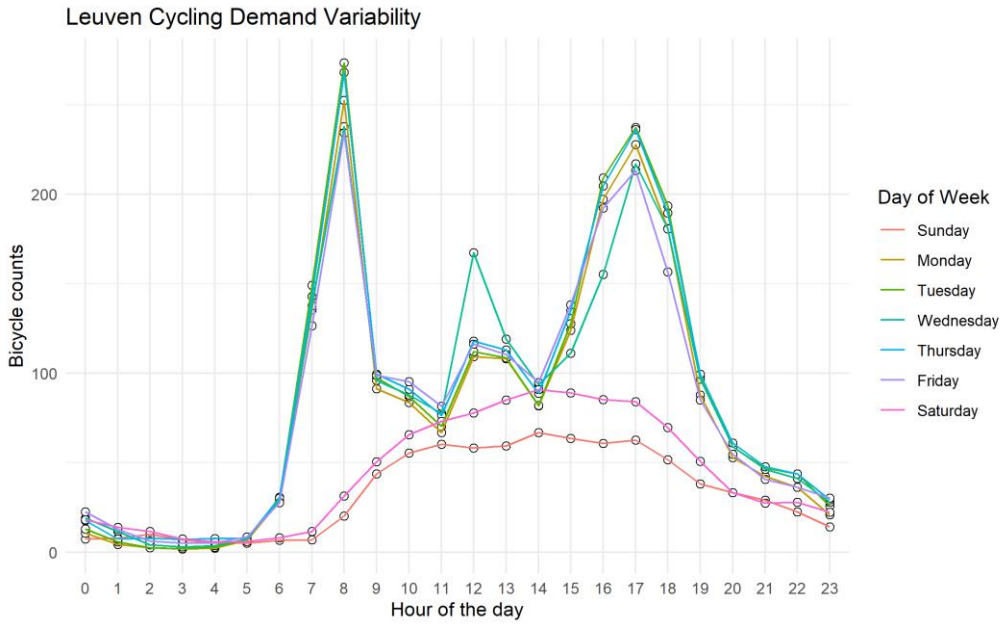


Figure 24 Leuven Case Study: Cycling Demand Temporal Variability

10.4.3 Pedestrian Demand

This section refers to the analysis of the data source described in Section 5.5 for pedestrian demand in Leuven, concretely the pedestrian data provided by Telraam. In this case, we performed completeness, validity and exploratory analysis with data from six randomly chosen sensors in the last six months of 2019.

10.4.3.1 Completeness and Validity Analysis

The analysis of this dataset concluded that the %MIV is 0, that is, there are no missing data on those time slots with at least some available data. Regarding the CPwMD, the figure below shows the distribution of periods without data. We can see that there are only two periods longer than 100 hours and that there is a peak in 15 hours. This peak is due to the characteristics of this data source since the Telraam devices only work during the daytime.

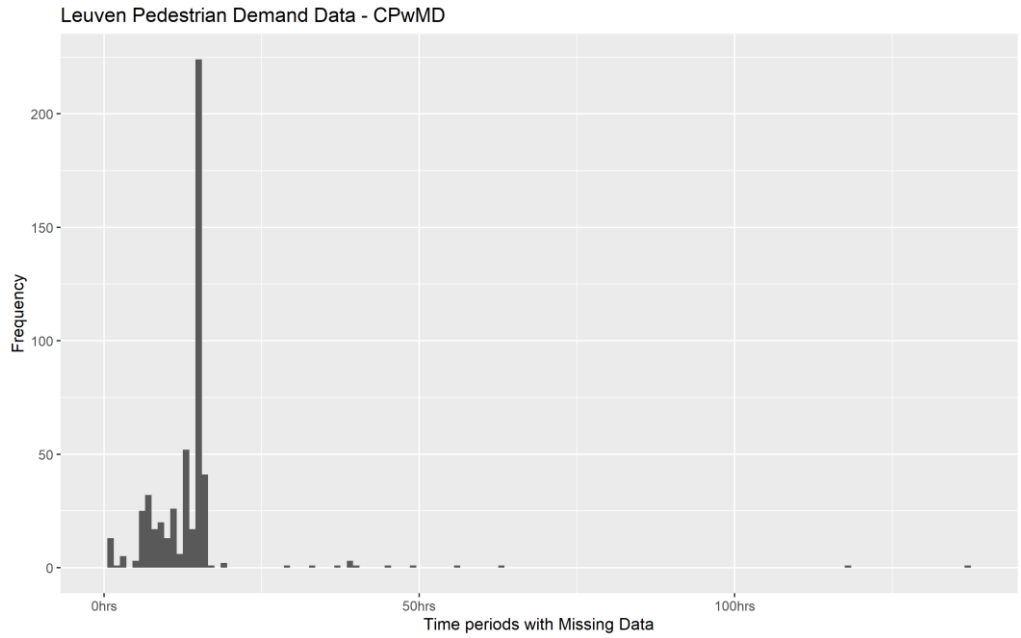


Figure 25 Leuven Case Study: Pedestrian Demand -CPwMD

10.4.3.2 Exploratory Data Analysis

The pedestrian demand variability per hour of the day and day of the week presents a different pattern than that of cycling. In this case, we cannot observe morning and evening peak times but two peaks at 12:00 pm and 4:00 pm on weekdays, respectively (see figure below). Furthermore, on Saturdays, the activity is even higher than on labour days.

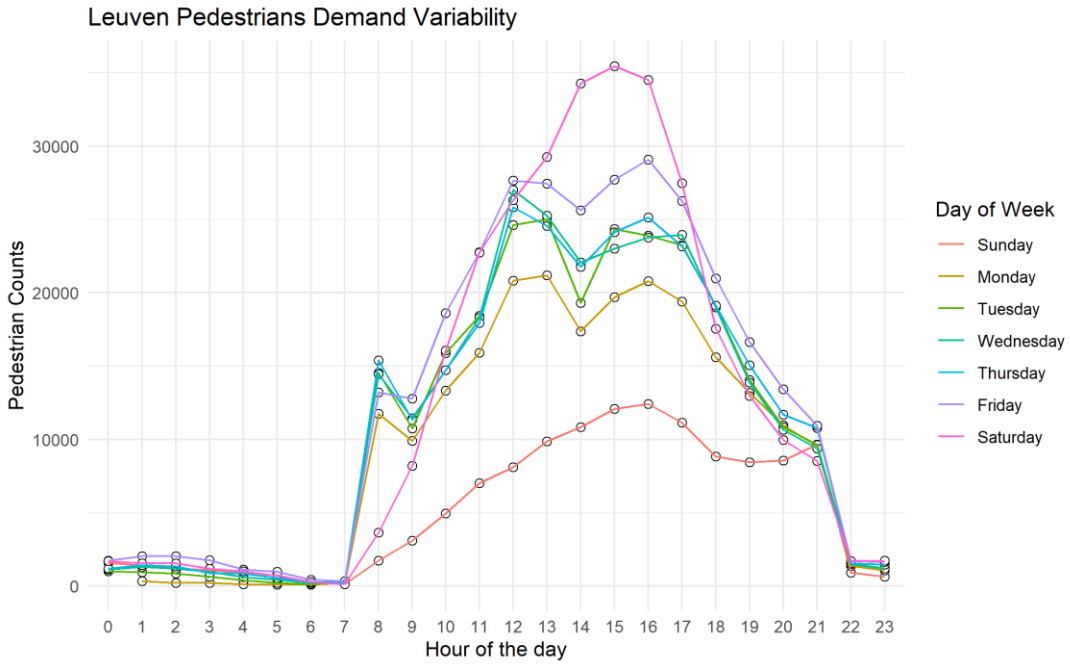


Figure 26 Leuven Case Study: Pedestrian Demand Temporal Variability

10.4.4 Traffic Data

This section refers to the analysis of the traffic data source for Leuven described in Section 5.9, concretely the vehicle detections also provided by Telraam. The sensors and analysis period are the same as those mentioned in the previous subsection.

10.4.4.1 Completeness and Validity Analysis

The results are the same as those discussed in Section 10.4.3.1, as measurements are the same, except that in this case, we consider vehicle counts instead of pedestrian counts.

10.4.4.2 Exploratory Data Analysis

The traffic demand variability per hour of the day and day of the week shows patterns more related to commuters with a morning peak from 7 to 9 am, and an evening peak at 6 pm, as it can be observed below. Traffic during weekends is also noticeably less.

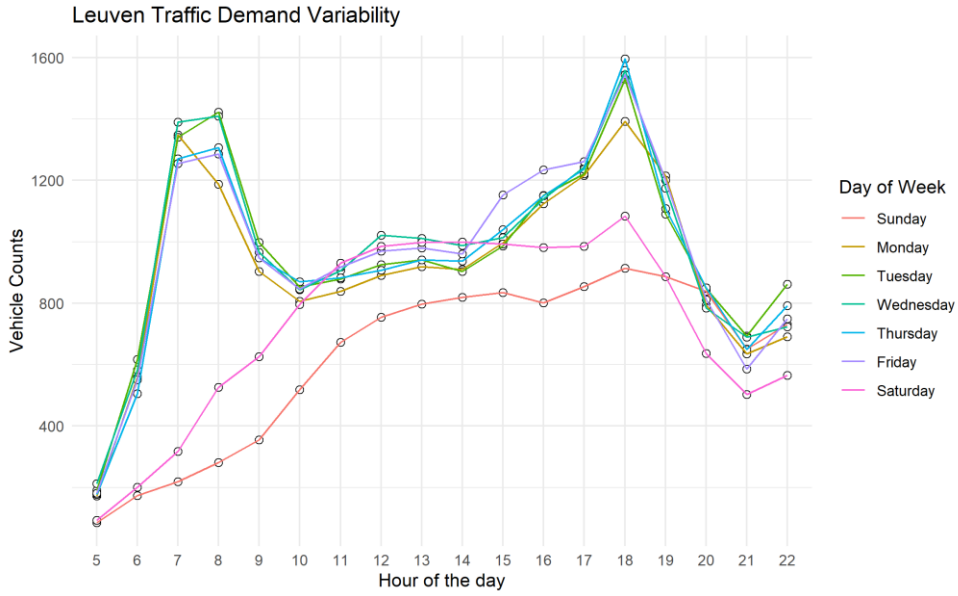


Figure 27 Leuven Case Study: Traffic Demand Temporal Variability

10.4.5 Parking Demand

This section refers to the analysis of the parking demand data source described in 5.12 for Leuven. In this case, we performed completeness, validity and exploratory analysis with data from the three available parking places with historical data from January 1st, 2018 until April 30th, 2019.

10.4.5.1 Completeness and Validity Analysis

The %MIV in this dataset is 0%, so there are no missing data on those time slots with at least some available data. Regarding the CPwMD, the figures below show the distribution of periods without data for each of the three parking facilities considered. It can be observed that there are only a few sporadic periods of 100 hours or more, and most are concentrated around periods of 5 hours or less. Besides, most of them occur at night, when there is no activity in the parking facilities. Therefore, the data source does not present relevant problems of completeness and validity.

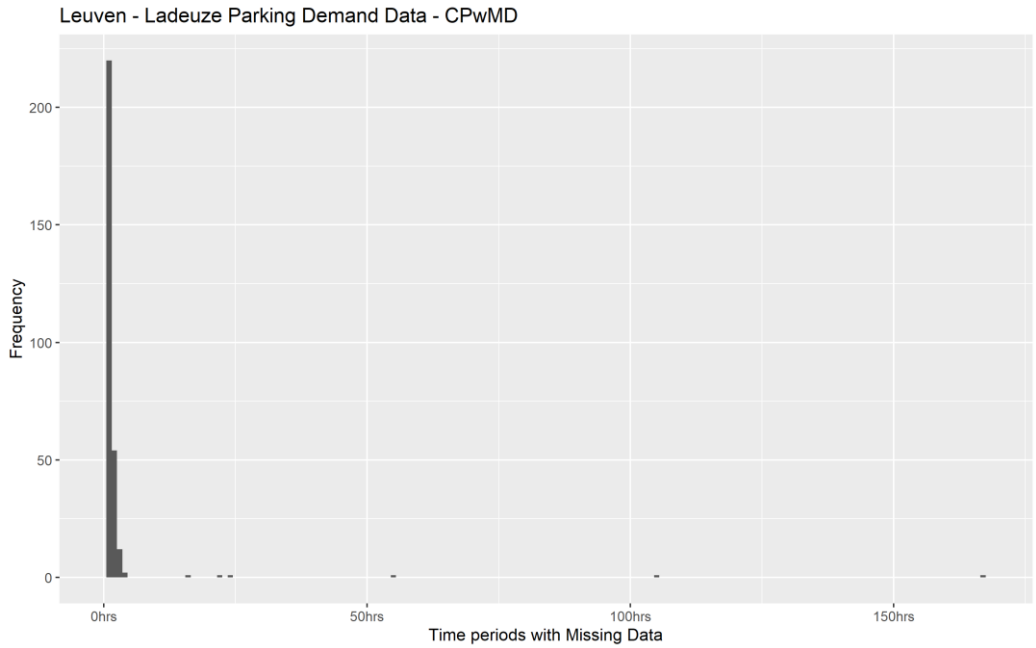


Figure 28 Leuven Case Study: Parking Demand – CPwMD I

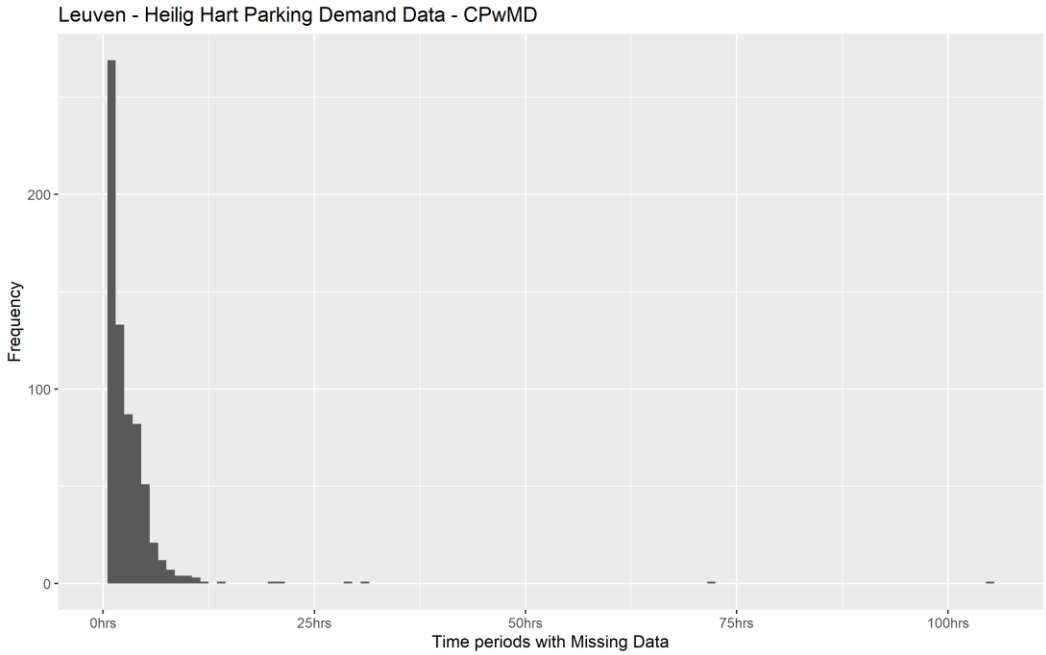


Figure 29 Leuven Case Study: Parking Demand – CPwM II

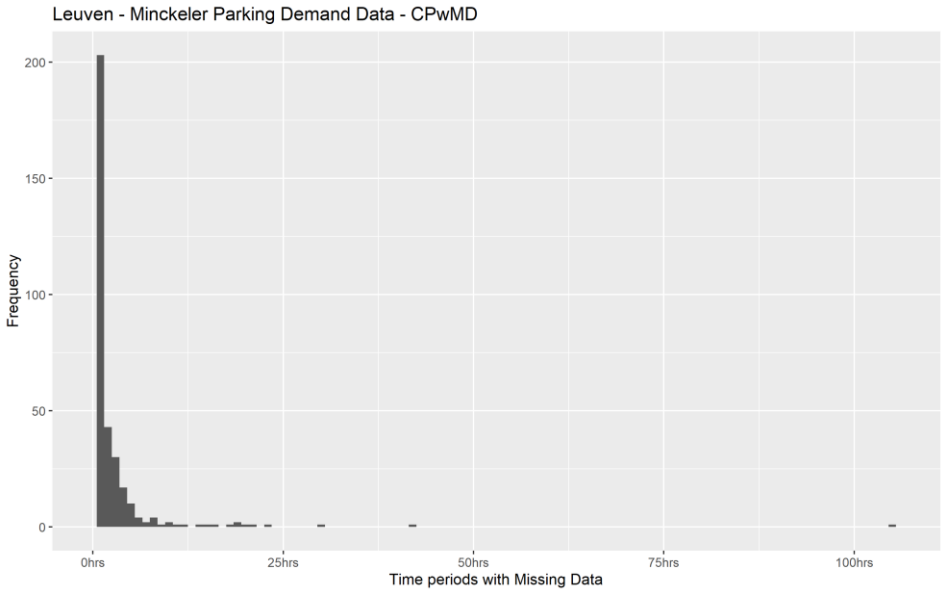


Figure 30 Leuven Case Study: Parking Demand – CPwM III

10.4.5.2 Exploratory Data Analysis

The temporal parking demand variability per hour of the day and day of the week is displayed in the figure below. In this case, we observe only one occupancy peak at around 1 pm on weekdays and 2 pm on Saturdays and Sundays. Interestingly, the occupancy on Saturdays is similar to that of weekdays, although delayed from one to two hours.

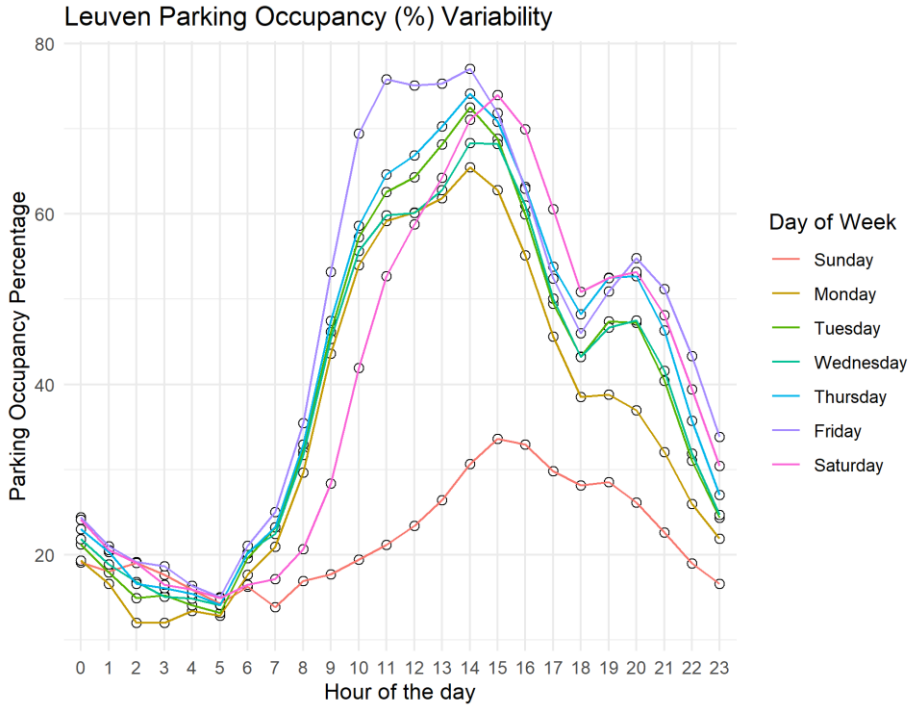


Figure 31 Leuven Case Study: Parking Demand Temporal Variability

10.5 Thessaloniki Case Study

10.5.1 Bike-Sharing Demand

This section refers to the analysis of the data source described in Section 5.3 for Thessaloniki, concretely the bike-sharing data provided by Thessbike. In this case, we performed completeness, validity and exploratory analysis with data from September 1st, 2019 till November 30th, 2019.

10.5.1.1 Completeness and Validity Analysis

The %MIV and CPwMD analyses concluded that there are no missing data nor periods without data in this case.

10.5.1.2 Exploratory Data Analysis

The bike-sharing demand variability per hour of the day and day of the week shows the frequency of use is low as usually there are no more than four trips per hour (see figure below). We can also observe that there are high differences between days of the week, although some general trends can be distinguished. In weekdays there are two peaks, one from 8 am to 9 am, and another one from 6 pm till 7 pm (except Thursdays). Sunday is the day with higher demand.

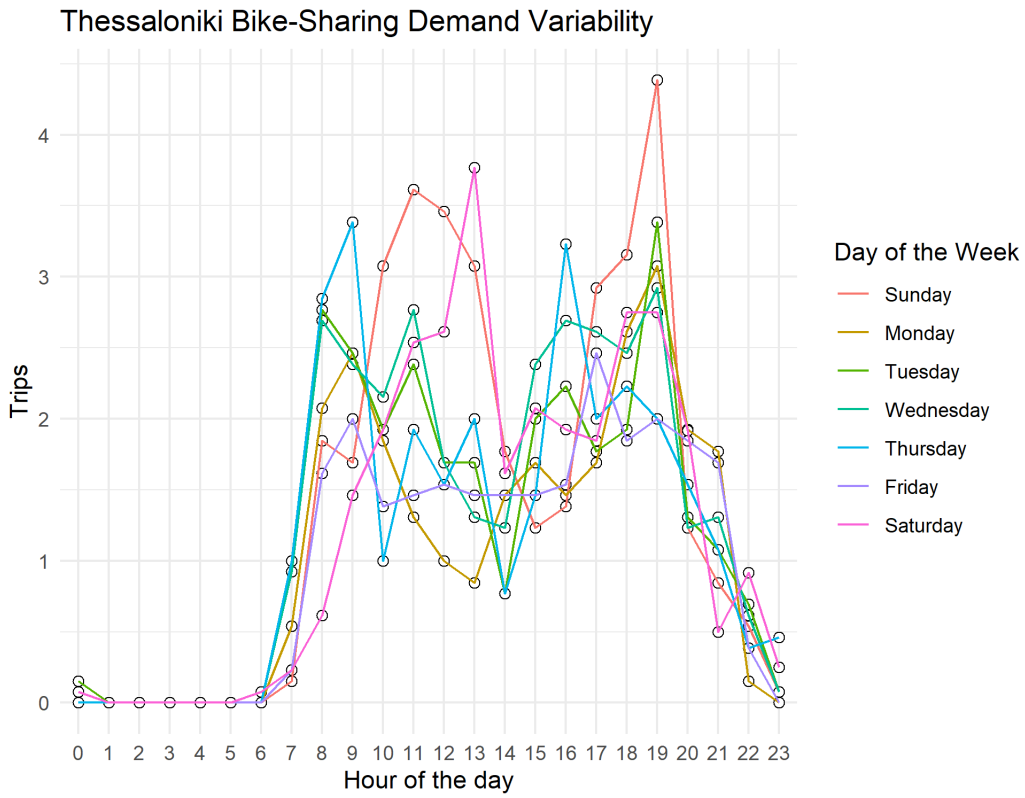


Figure 32 Thessaloniki Case Study: Bike-sharing Demand Temporal Variability

The figure below shows the trip duration distribution for Thessbike bike-sharing service. We can see that an important part of the trips has a duration between 15 and 30 minutes, although there is a relevant number of rides between 30 and 60 minutes.

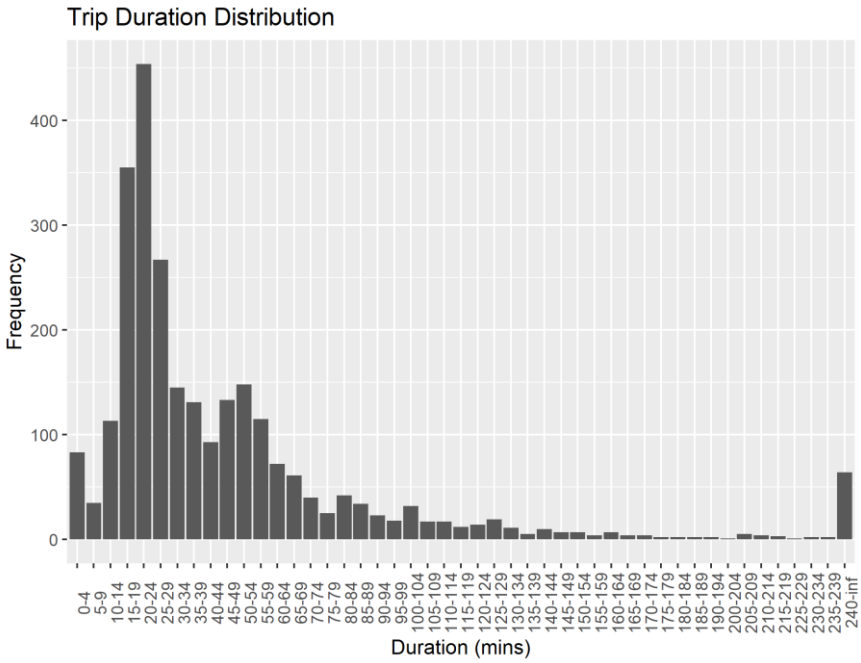


Figure 33 Thessaloniki Case Study: Bike-sharing Trip Distance Distribution

10.5.2 Traffic Data

This section refers to the analysis of the traffic data source described in Section 5.9 for Thessaloniki. In this case, we performed completeness, validity and exploratory analysis with data from September 1st, 2019 till November 30th, 2019.

10.5.2.1 Completeness and Validity Analysis

Due to the different nature of this data source, the completeness and validity analysis, in this case, was different in order not to obtain erroneous conclusions, and concretely, we only focused on validity. For the three months of interest, we noticed that 0.3% of records fall outside of the wider region of Thessaloniki and have either zero or too large values. Speed values (orientation and magnitude) are within expected ranges, speed is between [0,160] km/h and orientation that is an angle between [0,360).

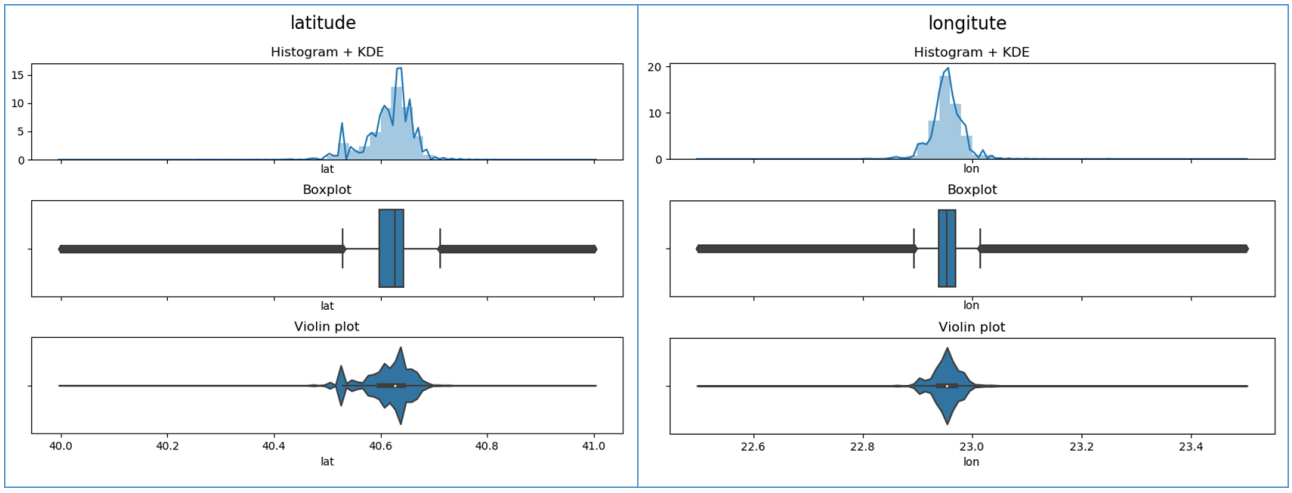


Figure 34 Thessaloniki Case Study: Validity Analysis of Traffic Data

10.5.2.2 Exploratory Data Analysis

The following figure shows the temporal variability of the percentage of network coverage, that is, the rate of street segments from the OSM map of the area of Thessaloniki through which at least one monitored vehicle passed. We can see that during the daytime, more than forty per cent of the network is covered in weekdays, where this percentage drops to 37 and 34 on Saturday and Sunday, respectively. In this way, we can conclude that the coverage of Thessaloniki's road network with this traffic data source is appropriate. Figure 36 shows traffic congestion levels visualized at three same-day intervals (morning peak-time, afternoon peak-time and valley period) estimated from this data source.

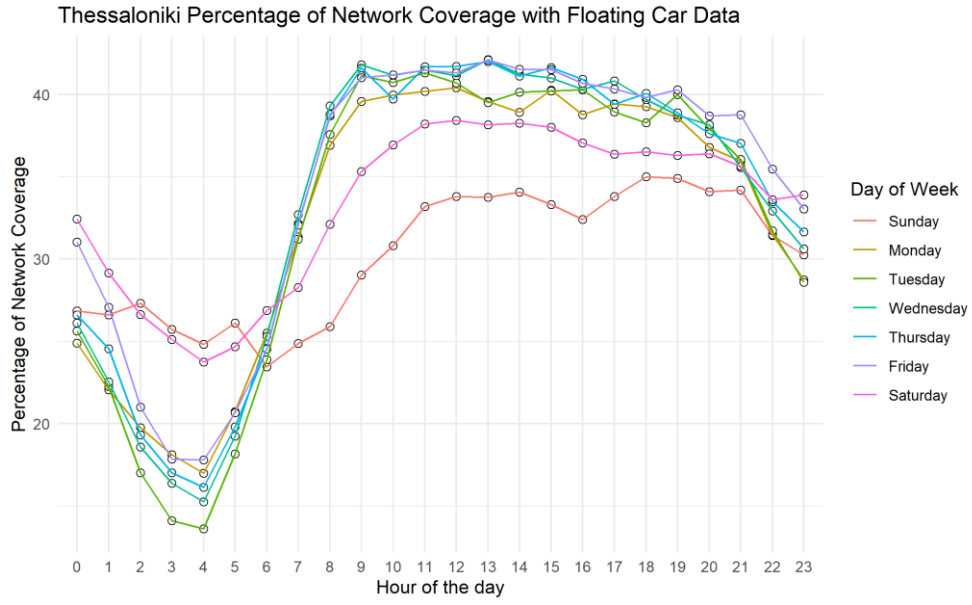


Figure 35 Thessaloniki Case Study: temporal variability of number of vehicles reporting floating data

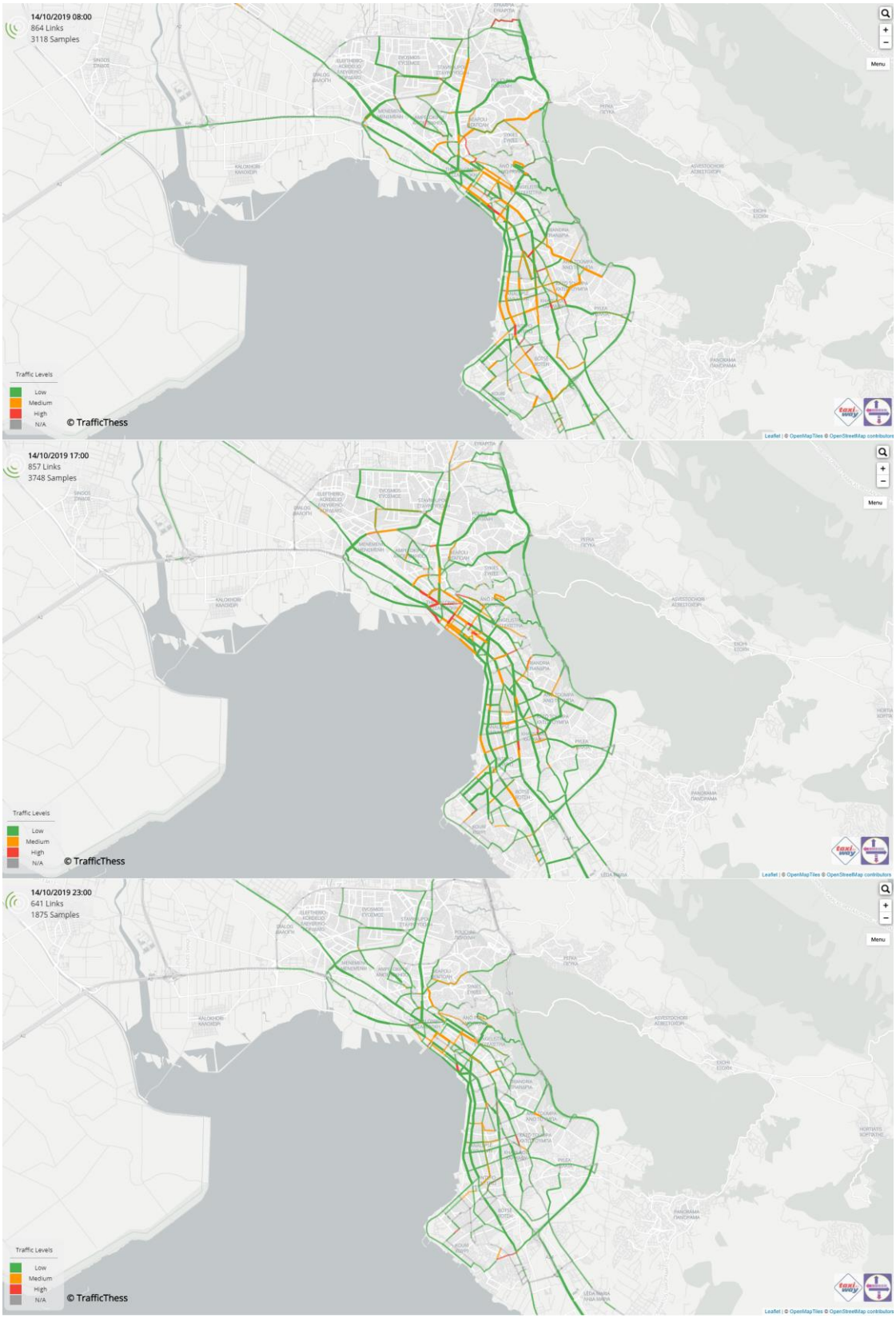


Figure 36 Thessaloniki Case Study: traffic congestion levels visualized at three same-day intervals (morning peak-time, afternoon peak-time and valley period)

10.5.3 Taxi Service Demand

This section refers to the analysis of the taxi demand data source described in Section 5.10 for Thessaloniki. In this case, we performed completeness, validity and exploratory analysis with data from September 1st, 2019 till November 30th, 2019.

10.5.3.1 Completeness and Validity Analysis

The %MIV and CPwMD analyses concluded that there are no missing data nor periods without data in this case.

10.5.3.2 Exploratory Data Analysis

The taxi service demand variability per hour of the day and day of the week, displayed in the figure below, shows two peaks in weekdays, one in the morning between 9 am to 12 pm, and another one in the evening between 5 pm and 6 pm. Analysing days of the week, we observe similar behaviour on Tuesday, Wednesday and Thursday. Monday show the same trend but with less demand in peak times. Sundays and Saturdays present a higher demand during the night but much lower during the daytime.

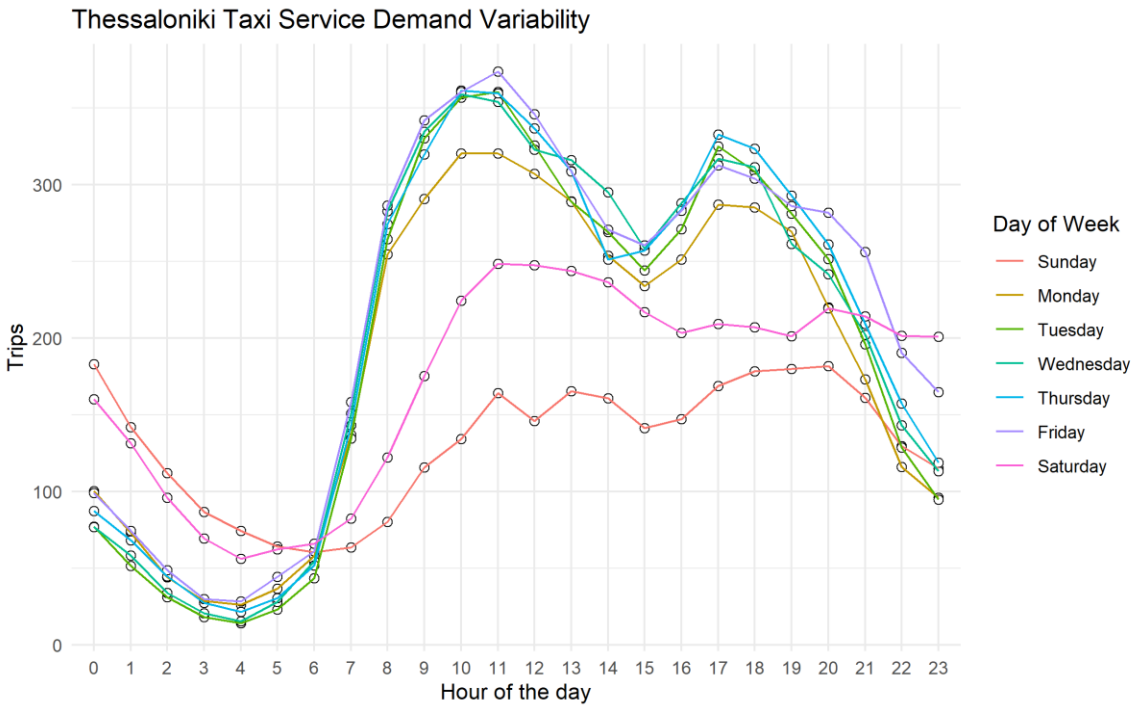


Figure 37 Thessaloniki Case Study: Taxi Service Demand Temporal Variability

The figure below shows the trip distance distribution for taxi service in Thessaloniki. We can see that the lengths of the most frequent trips are from one to four kilometres.

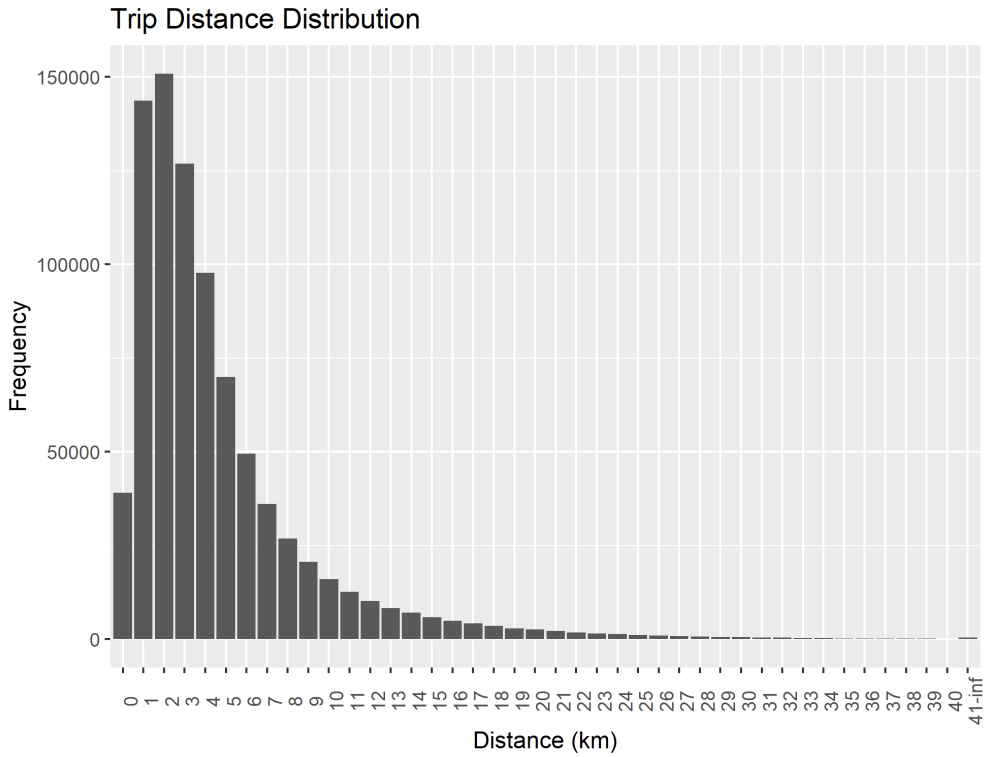


Figure 38 Thessaloniki Case Study: Taxi Service Trip Distance Distribution

10.5.4 Social Media Data

This section refers to the analysis of the Social Media data source described in Section 5.11 for Thessaloniki. In this case, we performed completeness, validity and exploratory analysis with data from September 1st, 2019 till November 30th, 2019.

10.5.4.1 Completeness and Validity Analysis

The %MIV and CPwMD analyses concluded that there are no missing data nor periods without data in this case.

10.5.4.2 Exploratory Data Analysis

The temporal variability of check-in events per hour of the day and day of the week is displayed in the following figure. Please, note here that we are considering three hours intervals. Interestingly, we observe a higher number of checking during the afternoon, evening and first hours of the night, which it is probably due to the bias of this type of data sources towards discretionary and leisure activities. This is also confirmed by the higher number of check-ins on the weekend. In any case, we can observe that the number of events collected by this data source is high, ranging on average from 500 to 1000 per three-hours interval on weekdays and from 1000 to 1500 per three-hours interval on weekends. In this way, we can confirm that the relevance of this data source for MOMENTUM is high.

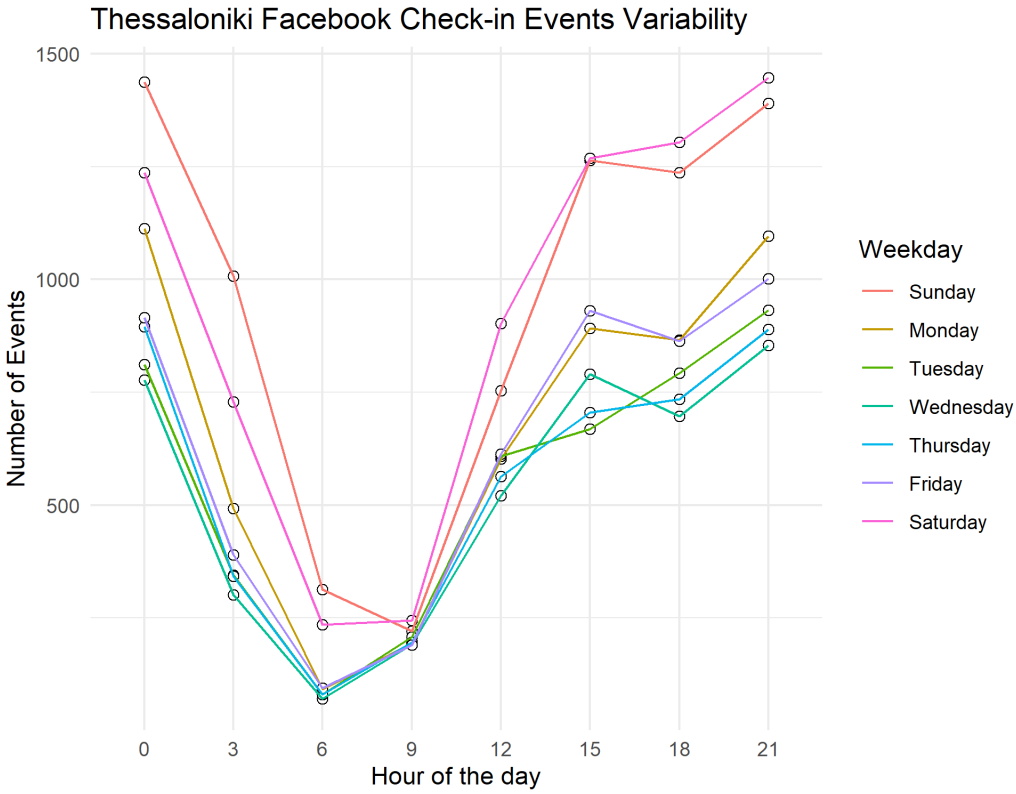


Figure 39 Thessaloniki Case Study: Facebook Check-in Events Temporal Variability