TANGENT

IN-DEPTH ANALYSIS OF TRAVEL BEHAVIOUR

D3.3

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 955273

Deliverable administrative information

Deliverable number	D3.3
Deliverable title	In-depth analysis of travel behaviour
Dissemination level	Public
Submission deadline	31/12/2023
Version number	1
Authors	Marios Giouroukelis (NTUA) Eleni I. Vlahogianni (NTUA)
Internal reviewers	Mark Brackstone (Aimsun)
Document approval	-

Legal Disclaimer

TANGENT is co-funded by the European Commission, Horizon 2020 research and innovation programme under grant agreement No. 955273 (Innovation Action). The information and views set out in this deliverable are those of the author(s) and do not necessarily reflect the official opinion of the European Union. The information in this document is provided “as is”, and no guarantee or warranty is given that the information is fit for any specific purpose. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein. The TANGENT Consortium members shall have no liability for damages of any kind including without limitation direct, special, indirect, or consequential damages that may result from the use of these materials subject to any liability which is mandatory due to applicable law.

https://twitter.com/TANGENT_H2020

https://www.linkedin.com/company/tangent-project/

https://www.youtube.com/channel/UCjhz4kwEm_sTHj7fE4zXToA

For further information please visit http://www.tangent-h2020.eu/

Executive summary

The goal of D3.3 is to conduct a comprehensive study on how introducing new modes of transportation and traffic management strategies affects the appeal of various transportation options and the resulting choices of commuters in selecting these modes. The key strength of this work is the use of two complementary data sources that allow for the creation of econometric mode choice models that integrate both commuters' perceptions on the attractivity of different modes, as reflected in stated preference data, and their real-world choices, as indicated by revealed preference data of their actual trips.

For the development of mode choice models, the study examined two distinct scenarios for Athens. The first scenario involved a choice between using a car and public transport, with a focus on prioritizing public transport over cars. The second scenario presented a choice among using a car and paying a toll for entrance to the city centre, using a car and bypassing the area where the congestion pricing scheme is applied, and using public transport.

The stated preference data were collected through TANGENT's travel behaviour survey, which centred on stated behaviour scenarios. In these scenarios, participants were presented with hypothetical situations and were asked to select the most attractive mode alternative. Each questionnaire also documented the users' mobility profiles, their perceptions towards new services and traffic management strategies, as well as their sociodemographic characteristics.

The SP data were subsequently enriched with RP data, reflecting commuters' actual mode choice decisions made in real-life situations. The RP data were collected from participants who provided trip trajectory data tracked by the Google Maps app on their smartphones over a month-long period. By analysing the transportation modes that commuters selected between various starting points and destinations, and by modelling the attributes of the modes they did not choose, we were able to reconstruct the users' real travel decisions.

Both datasets were then combined into a joint SP-RP database. This approach allows for the exploitation of the advantages each data source offers, while also facilitating the correction of the inherent biases present in each dataset. The particularities of the fusion of different data sources and the creation of joint SP-RP models are extensively discussed, addressing various methodological issues.

To identify the most influential factors in mode choice, econometric multinomial logit models were trained and tested on the joint databases. These models were specifically designed to account for the fact that the data came from different sources, incorporating a scale factor to adjust variance differences between SP and RP data. The models' predictive capabilities were evaluated using the accuracy, precision, recall, and F1-score metrics, showing strong performance for both scenarios.

After the estimation of the models, interesting insights are drawn regarding the factors that most significantly influence transportation mode choice. Specifically, increases in travel time and cost tend to reduce the perceived utility of every mode. Additionally, factors such as an individual's socioeconomic profile, including occupation, income, and gender, as well as preferences and perceptions of each mode's quality and commuter reaction to unexpected events impact travel choices significantly.

The study highlights the potential of accurate predictions of the modal split to assist policymakers and transportation in the promotion of sustainable transportation modes. By estimating the impact of new services and traffic management strategies on travel demand for different modes, these predictions

can help identify factors influencing commuter behaviour and enable the design of effective strategies accordingly.

Key words

Travel Choice Modelling, Discrete Choice Analysis, Stated Preferences, Revealed Preferences, Joint SP-RP modelling

Table of contents

List of figures

List of tables

List of abbreviations and acronyms

Acronym	Meaning
TBS	Travel Behaviour Survey
SP	Stated Preferences
RP	Revealed Preferences
WP	Work Package
MNL	Multinomial Logit

1. Introduction

In the realms of transportation research and urban planning, understanding travel behaviour is essential, especially in cases where multi-modal and automated travel options become available, new traffic management measures are applied in the network, or when unexpected events occur. It has been widely discussed (see D3.1) that the availability of new modes, or the change in attractivity of the existing ones can significantly alter mode choice patterns. Consequently, there's a need to evaluate retrospectively the impact of these new contingencies. This assessment is vital for both government officials, in shaping policy, and for the private sector, where understanding travel behaviour is key to informed decision-making and successful transport-related product launches.

Typically, to model the impact on mode choice of strategies or modes not yet present in the market, researchers conduct stated preference surveys. These surveys gather commuters' preferences by asking them to react to hypothetical scenarios. While widely used and with proven econometric merits, there always remain the question as to which extent the stated preferences are able to reflect the actual commuters' choices.

In this study, we are interested in merging data on mode choice preferences, as stated by individuals, with their actual mode choices, as revealed in real-life situations. To collect the revealed preference data, selected commuters were requested to share with the TANGENT team data of their actual trips conducted within a period of one month. The analysis of the collected trips enables the reconstruction of mode choice decisions as they occur in a real-world setting. Utilizing the joint SP-RP dataset, in-depth econometric models can be developed, offering valuable insights into the factors influencing individuals' mode choice patterns.

1.1 Attainment of the objectives and explanation of deviations

As per the Grant Agreement (GA), there are two objectives of WP3 that should be addressed and are relevant for this deliverable:

Objective B: Development of a set of models to describe travel choices and mobility shift in both an aggregated and in-depth level.

This objective is addressed by the development of econometric models for the description and quantification of travel choices and mobility shifts. The models have been trained on a joint database, comprised of both stated and revealed preference data. By combining these two types of data, the developed models can provide a more comprehensive understanding of commuter behaviour, capturing both hypothetical inclinations and real-world actions with respect to mode choice.

Objective C: Identification of factors that affect behavioural shift under various traffic management measures (strategies) and unexpected events.

These issues have been addressed through the SP survey design, the analysis of the revealed commuter's mobility patterns and in the incorporation of both data types in the developed models. Commuters' reaction to the network impact caused by each of TANGENT's traffic management strategies are approached by specific survey questions. The stated behaviour of commuters is supplemented by insights revealed from their actual mobility patterns. These real-life commuting patterns provide a concrete context to their hypothetical responses, revealing more realistic

preferences with respect to mode choice. In-depth econometric models are trained on the joint SP-RP data, capturing successfully the factors affecting behavioural shift.

Overall, the objectives related to this deliverable have been achieved in full and as scheduled.

1.2 Intended audience

This deliverable is public, thus accessible not only to TANGENT partners and project officers, but to anyone interested in travel behaviour analysis, mobility patterns and travel behaviour shifts that may occur when new traffic management measures are implemented in the road network.

1.3 Structure of the deliverable and links with other work packages/deliverables

The deliverable builds upon the literature review in mode choice modelling covered in D3.1 and advances on the mode choice models that are solely based on stated preference data, presented in D3.2. For the development of travel choice models, input was required from the case studies (WP7) to further define which TANGENT services and traffic management strategies are relevant to each pilot, leading to the development of city-specific versions of the travel choice questionnaires. In addition, both the survey and models address the particularities (from a mode choice perspective) of the traffic management strategies developed within WP5, described in detail in Deliverable 5.2 “Optimization Models for Transport Network Management”. The models presented in this deliverable will provide demand forecasts regarding the number of commuters who will use either a car or public transport when TANGENT's strategies are implemented. This information is necessary for the testing of WP5 algorithms within a simulation setting such as the virtual case study of Athens (WP7).

This document outlines the process and methodology for creating travel choice models using a combination of stated and revealed data. Chapter 2 provides a short literature review of recent methodological developments and applications of joint stated and revealed preferences modelling. Building on this theoretical background, Chapter 3 presents the methodological approach for training and testing discrete choice models on joint SP-RP databases. Additionally, the collection and analysis of revealed preference data based on actual user trips are extensively discussed. Chapter 4 outlines the implementation of the models within the TANGENT solution framework and discusses the various factors that affect mode choice. Chapter 5 concludes the deliverable.

2. Stated and Revealed preference Mode Choice

2.1 Literature Review

Mode choice models play a major role in transport demand modelling providing a key input for officials and engineers in the effort of directing infrastructure investments towards a more efficient and user-friendly transportation system (Tabasi et al., 2023).

Stated Preference (SP) data have resisted the test of time, providing the necessary input for the development of mode choice models in various research contexts since the very beginning of mode choice modelling (Small, 1987). A major reason for this is the relative ease with which one can collect the necessary datasets using surveys. SP surveys quantify how people might behave in a situation by presenting them structured “scenarios” where they are asked to choose between different policies, products, services, or, in the context of travel choice modelling, modes, having both desirable and undesirable attributes. These controlled experimental conditions allow for the introduction of sufficient attribute trade-off variability and reduction of collinearity in data, which facilitates their analysis during their statistical modelling (Louviere et al., 2000). Often, relying to an SP survey is the only way to generate data, especially when considering hypothetical scenarios and alternatives that may not be feasible or existing in an actual urban setting (Tabasi et al., 2023). Common examples are situations where new transit modes are being introduced, and there is no existing data to analyse their impact or whenever of public goods that do not carry evident price signals (such as environmental amenities) need to be estimated (Yan et al., 2019). Even in the case of highly abstract scenarios, the numerous applications relying on SP data prove that their use is quite effective: respondents typically have a solid understanding of hypothetical choices and can make reasonable trade-offs among attributes (Louviere et al., 2000)

While conducting SP choice experiments is a common practice, their reliability in reflecting real market behaviour remains questionable. The most common critique is their susceptibility to various biases: the context and format of the hypothetical scenarios can influence responses, often leading to an artificial setting (Hanley & Czajkowski, 2019). On that matter, some studies that have observed disparities between stated choices and real market behaviour further contribute to the criticism of SP data's reliability (Morikawa, 1994). For these reasons, economists have a long-standing tradition of favouring Revealed Preferences (RP) data, relying on the idea that commuter preferences can be better deduced from their actual behaviour (Börjesson, 2008). Drawing from real decisions, RP data eliminate the need for assumptions about how consumers might behave in hypothetical scenarios, and their inherent connection to real-world choices is considered more trustworthy than SP preferences (Mark & Swait, 2004). Consequently, recent literature has shown a growing interest in using RP data for various mode choice modelling (Yan et al., 2019; Ilahi et al., 2021; Lizana et al., 2021;). To elicit RP data, researchers often rely on commuters’ self-reported information about their trips, but this method has its drawbacks, including potential limitations such as underreported trips and reduced accuracy (Tabasi et al., 2023). In recent times, there has been a shift towards utilizing GPS-based smartphone applications to comprehensively track commuter trip choices, surpassing the capabilities of traditional travel survey data (Huang et al., 2021).

Consequently, a key question is how one can profit from the unique strengths of both data types while mitigating their individual limitations. The recent literature on the integration of Stated Preference (SP) and Revealed Preference (RP), also referred to as 'data fusion' or 'data enrichment,' marks a pivotal advancement in the fields of consumer behaviour analysis and transportation research, as it has been shown that SP data can complement RP data (Lavasani et al., 2017). The key intuition for this data fusion is that, when making decisions, individuals tend to value similarly the attributes (e.g. travel time,

cost etc.) of each mode choice, as well as the trade-offs between them, regardless of whether the decisions are made in a SP or RP setting.

The combination of RP and SP data offers several key advantages. Firstly, it enhances efficiency through the joint estimation of a model's parameters, leveraging the comprehensive data available from both sources. Secondly, it allows for more realistic modelling results by incorporating real-world information from RP data, thereby reducing response biases inherent in SP data. Thirdly, this approach enables the identification of preferences for new products, services, or attributes that are not discernible from RP data alone (Ben-Akiva et al., 1994). Additionally, by combining both data sets, researchers can improve the variability-attribute trade-offs that may be underrepresented in RP data, because of their "unstructured" nature. Intuitively, it can be considered that RP data serve as the "ground truth" regarding commutes' choices, while SP data enriches them by introducing additional information, such as variability in each modes' characteristics, or additional choices that have not yet been introduced to the market (Yan et al., 2019).

The practice of applying discrete choice models, has become a standard approach in the study of combined SP-RP databases, owing to their flexible structure that accommodates some particularities of joint databases, such as their differences in the relative variance between them (Lavasani et al., 2017). Notably, their use in applications involving the introduction of traffic management strategy such as congestion pricing schemes is especially relevant, as it aligns closely with the objectives of TANGENT. After the application of Stockholm's congestion pricing scheme in 2006, (Börjesson, 2008) analysed trip timing decisions by applying a mixed logit model on joined SP-RP choice data. Her methodology combined SP scenarios eliciting commuters' preferred departure and arrival time with RP trip information derived from roadside number plate registrations from traffic cameras. The RP trips have been further annotated by commuters themselves, who have provided information on their origin, destination, arrival time etc. In a similar application, (Lizana et al., 2021) analysed the impact of a future congestion pricing scheme in Santiago, Chile on travel behaviour, focusing on trip departure time decisions. The SP data is derived from a series of structured scenario experiments, while the RP data consist of actual trip information collected for a specific day in 2021 inputted in a trip diary along with socio-demographic data and employment characteristics of the respondents. Afterwards, joint mode-departure time choice models with various specifications, including Multinomial Logit, Nested Logit, and Mixed Multinomial Logit models were developed.

A second relevant field of applications aim at the ex-ante estimation of changes in mode choice patterns caused by the implementation of new transport modes. Studying mode choice decisions at the integration of ride sourcing services (such as on-demand, app-driven ride-sharing services) with public transportation, (Yan et al., 2019) has proposed a series of mixed logit models trained on a joint SP-RP database. For collecting the RP data, respondents replied to a questionnaire survey on their current mode choices and trip attributes. Specifically, they were asked to identify their most frequently used mode of transportation from a list of six alternatives and then provide estimates of travel attributes associated with the four main travel modes. Similarly, (Ilahi et al., 2021) proposed various multinomial and mixed mode choice models to estimate demand for novel future on-demand transport modes such as Urban Air Mobility (UAM) solutions. In addition to the questionnaire-based SP scenarios, the RP data have been elicited using a travel diary spanning three working days, documenting respondents' travel behaviours, including departure times, transport modes used, destinations, and the purposes of their trips.

2.2 Methodological approach

The main goal of this deliverable is to develop accurate discrete choice models that can describe how travel mode decisions are made, by combining both stated and revealed preferences data sources.

Stated Preference (SP) data were gathered using the questionnaire-based Travel Behaviour Survey (TBS). The survey collected diverse information, including sociodemographic details, travel habits, perceptions of transport system attributes, and preferences for various travel alternatives, presented in specific scenarios where different mobility strategies are implemented on a network. Tailored to each pilot city, site-specific questionnaires were created to address the unique mobility needs and network conditions of each location. Data collection was facilitated using the user-friendly interface of project partner Cefriel's CONEY toolkit. The interested reader is referred to D3.2 for an extensive presentation of the SP data collection process.

Revealed preference (RP) data were collected by analysing actual mode choices of a sample of commuters who provided data on their daily commute across a period of 30 consecutive days. Leveraging the trip tracking capabilities of modern smartphones, participants were guided into enabling Google Maps Timeline features on their devices, exporting their trip data when the collection of the necessary sample has been achieved, and sharing them with TANGENT researchers. The commuters who have shared trip data have also participated in the SP-leg of the survey.

Both phases of the data collection have been actively framed by WP3 technical consultants, as well as case study leaders in each of Athens, Lisbon, Manchester, Rennes.

After the data collection activities concluded, the collected data has been pre-processed and consolidated into two databases suitable for modelling. Of particular interest is the procedure that has been followed in the case of the RP data, where through analysing the routes and modes that commuters took, we successfully extracted data on travel time and cost for both their chosen and non-chosen alternatives. Leveraging the capabilities of the Google Maps API, the travel time costs associated with the non-chosen options, a detail often difficult to gauge or report accurately, was calculated with high fidelity.

After the pre-processing of both SP and RP datasets, the data are merged and the joint database is used in the creation and calibration of discrete choice models in two scenarios: "Prioritisation", where traffic control allows preferential crossing of public buses in signalized intersections and "Pricing", where a congestion pricing scheme is applied to disincentivise car trips to the city centre. For each of the scenarios, Multinomial Logit Models are developed and discussed regarding the accuracy on their predictions. Reflecting on their output, useful insights can be drawn on the impact of novel traffic management strategies on the modal split.

The above-described methodology from the collection of the data to the development of the models is depicted in Figure 1.

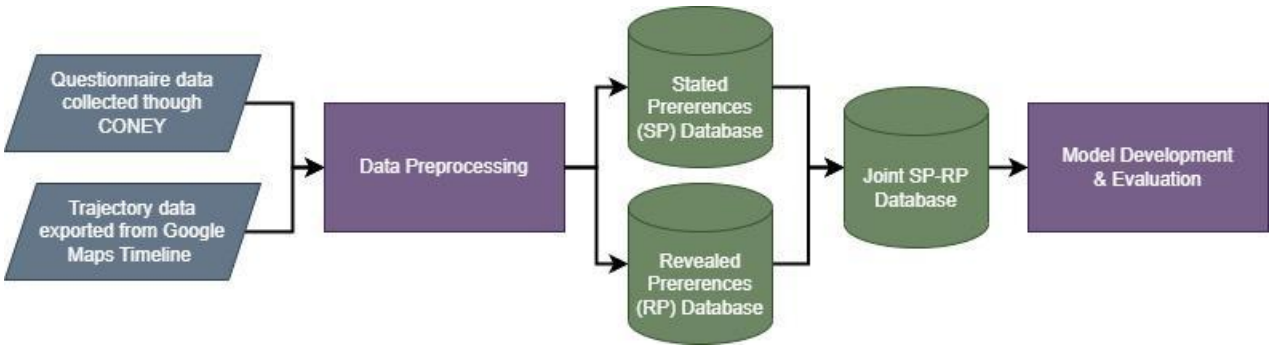


Figure 1: Travel choice modelling using SP & RP data pipeline

3. Data Collection and Pre-processing

In this section, a thorough description of the typology, collection and preprocessing of both stated and revealed preference data used in this study is provided and discussed.

3.1 Stated preferences

The stated preference data that have been used to calibrate the in-depth mode choice models were collected through the WP3 Travel Behaviour Survey. The aim of the questionnaire was to elicit commuters' perceived travel patterns that can be compared to their actual trips

To achieve this, the core of the questionnaire is formed by several stated preference scenarios, where respondents select their preferred mode for a trip in the city centre, by evaluating each mode's characteristics such as commuting time and cost. In here, we are reintroducing the stated preference scenarios, as it facilitates the discussion of Section 3.1.2.2 on how actual trip data collected from commuters can be used to recreate similar choice scenarios in an RP setting.

The "Synchronisation" scenario illustrates the case of a trip from the suburbs to the city centre, where individuals can choose between driving a private car and using public transportation. In this scenario, public transport systems (specifically buses in the case of TANGENT) are aligned with traffic control which increases the average speed of the vehicles. This scenario addresses the changes of the demand patterns when the algorithms developed for the *Synchronization of Public Transport and Traffic Control* optimization problem of Service 2 of TANGENT are applied in an actual or simulated city environment. The interested reader is referred to Deliverables 5.2 and 3.2 for a more thorough description of TANGENT's optimisation problems and how changes in demand are addressed. In the Athens-specific version of the questionnaire, the formulation of the scenario is as follows:

> ***In order to perform a typical trip from the suburbs to the centre of Athens, with length of approximately 10 km, during peak hours, the available modes are car and Public Transport. To improve traffic conditions traffic light management strategies that give priority to Public Transport are applied.***

The situation described above involves participants being given six hypothetical choice situations, where they are requested to choose between two alternatives. These alternatives vary in terms of transportation mode (either car or public transport), the cost involved, and the duration of the trip, including both travel and wait times. An illustrative example is provided in Table 1.

Table 1: One of the six choice situations for the "Synchronisation" scenario

Option 1 Mode: Car In-vehicle time: 65 minutes Waiting time: - Cost: 5€	**Option 2** Mode: Public Transport In-vehicle time: 55 minutes Waiting time: 15 minutes Cost: 2€
○	○

The "Pricing" scenario presents a situation akin to the first, where commuters can travel to the city centre using either a car or public transport. In this case, a congestion pricing scheme is implemented, increasing the overall cost of entering to the city centre by car. This configuration of the transport network provides car users with two choices: pay a toll to enter the city centre or avoid the toll area entirely. This scenario is linked to the *Dynamic Congestion Pricing* optimization problem of Service 2 of TANGENT and its description is as follows:

> ***"For a typical trip from the suburbs to a central neighbourhood of Athens during morning peak hours, commuters can choose between using their private car or opting for the public transport system. To improve traffic conditions, a congestion pricing scheme is applied: car users who wish to pass through the city centre must pay a toll. Commuters who are not willing to pay the toll but nevertheless want to continue using their car can bypass the tolled area and use peripheral roads to reach their destination."***

The scenario involves once again six hypothetical choice situations but this time there are three travel alternatives, differentiated with respect to the mode (car with toll payment, car without toll payment or public transport), cost, distance, and time (Table 2).

Table 2: A choice situation for the "Pricing" scenario

Option 1 Mode: Car, payment of toll Distance: 6.5 km Time: 30 minutes Cost: 3.5€	**Option 2** Mode: Car, bypass the tolled area Distance: 8.5 km Time: 35 minutes Cost: 2€	**Option 3** Mode: Public Transport Time: 45 minutes Cost: 0.60€
◯	◯	◯

In addition to the stated preference scenarios, each questionnaire includes questions on the mobility profile of commuters, their perception regarding different system specifications and new modes, their reactions to unexpected events, as well as their sociodemographic characteristics. A thorough description of the other parts of the questionnaire, its adaptation to each case study's context and the dissemination procedures linked to it is included in Deliverable 3.2. These details are not revisited in the current document.

3.2 Revealed preferences

Google Maps Timeline is a feature within Google Maps that allows users to view and manage their location history. This feature is tied to the Location History setting in a user's Google account, and it tracks places they've visited with their mobile devices where they are logged into their Google account (Figure 2). Specifically, Google Maps Timeline records the places users have visited, the routes taken, while it also provides an informed guess on the mode of transportation used, based on the device's location data and the application of advanced machine learning classification algorithms developed by

Google Maps Timeline. Extracting data from Google Maps Timeline is equally feasible, since they are kept stored in the user's Google account and can be exported and shared.

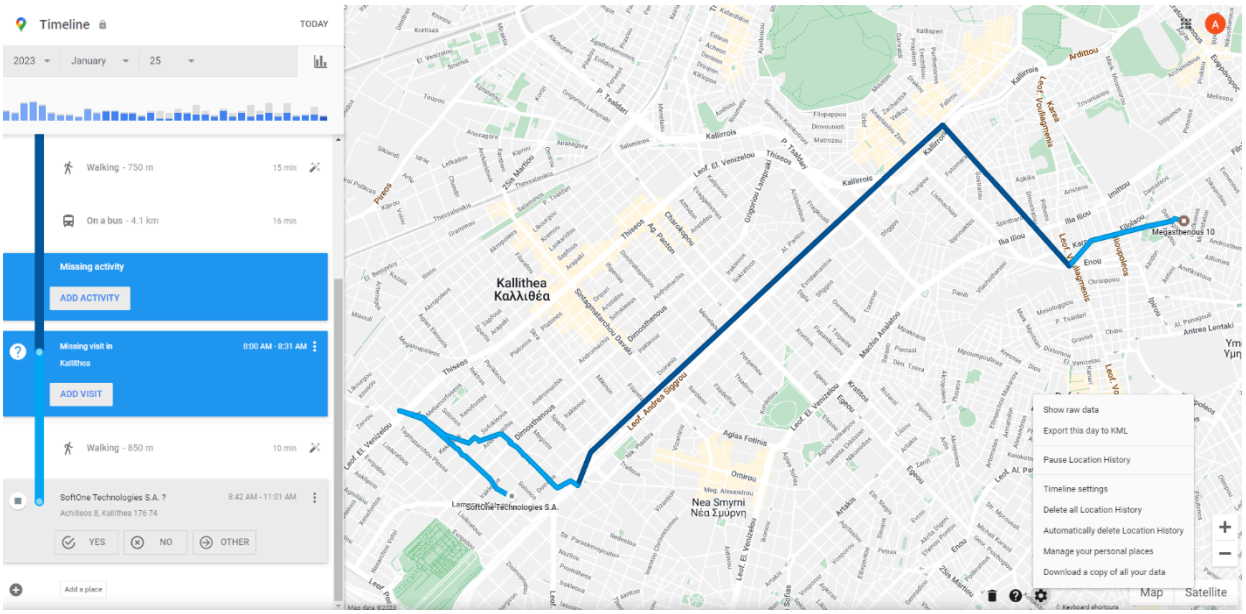


Figure 2: A typical route, as it appears on the Google Maps Timeline Dashboard

For the purposes of the research conducted within TANGENT, participants were asked to export and share with the WP3 team the trips of a typical month (30 days). This can be achieved by opening Google Maps on a web browser and sign in with the Google account that is used for tracking location history. Typically, this is the account that has been used to log in to Google Maps from the user's mobile device. Then, by navigating to the Timeline section and selecting a specific date, the user can view the trips and visited locations, and download them in a KML format. The KML file, short for Keyhole Markup Language, includes various features like place marks, images, polygons, 3D models, and textual descriptions. These features are then used by geospatial software that supports KML encoding to display them on maps. The structure of a simple KML document containing the coordinates of New York City is included in Figure 3.

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2">
<Document>
<Placemark>
  <name>New York City</name>
  <description>New York City</description>
  <Point>
    <coordinates>-74.006393,40.714172,0</coordinates>
  </Point>
</Placemark>
</Document>
</kml>
```

Figure 3: A KML file containing the coordinates of New York City

The extraction process is repeated for each day, for the entire span of a typical month. For each day, two files are exported: one containing a user's visited places, and the second one containing trajectories, which originate from or are directed towards these places (Figure 4).

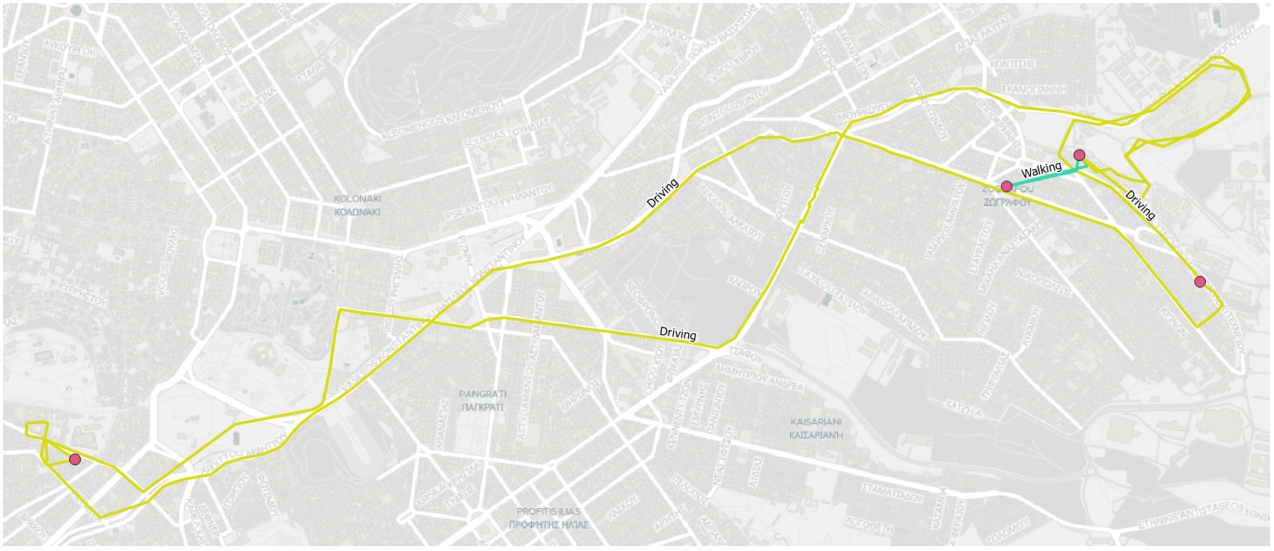


Figure 4: Visual description of a day of commuting for a participant in Athens

A staypoint is a location where an individual or vehicle remains stationary for a significant amount of time, as determined by GPS tracking data. In the context of GPS-based monitoring and analysis, a staypoint is identified when the GPS data shows that the subject has stayed within a certain small area for a predefined duration. This could indicate places like homes, workplaces, or any other location where someone spends a considerable amount of time. Staypoints are defined by their coordinates, which are numerical representations of their location in a spatial reference system. In addition to its coordinates, each staypoint is annotated by the name and the category (such as *building* or *copy shop*) of the visited location, and the duration of the stay (including a "beginning" and an "end" timestamp). An actual example is included in Table 3.

Table 3: Annotation of staypoints

Name	Begin	End	Category
Home	2022-11-30 20:25:50	2022-12-01 07:15:31	Building
Fasma Copy Chop	2022-12-01 07:54:53	2022-12-01 08:11:51	Copy shop
Work	2022-12-01 08:17:34	2022-12-01 10:53:15	University department
Roumeliotis	2022-12-01 10:58:30	2022-12-01 11:23:28	Pastries
Work	2022-12-01 11:27:44	2022-12-01 16:36:23	University department
Home	2022-12-01 17:17:41	2022-12-02 07:38:48	Building

The second file contains user trips as files of the type “LineString”, a type of geometry used in geographic information systems (GIS) to represent a linear feature composed of a sequence of points, connected by straight line segments. Each segment is equally annotated by its name (denoting in this context the mode that was used for the trip), the duration of the movement (again denoted by a “beginning and “end” timestamp) and the trip’s distance. The trip information which is associated with the origin and destination points of Table 3 are presented in Table 4.

It is also noted that both files also contain identifiers that permit the association of each trip and stay point to a user, which allows for the association of each user’s trips to her responses on the stated preference questionnaire.

For more information about the Google Maps timeline functionalities, the interested reader is referred to the Google Maps help portal (available here for the readers of the online version of this deliverable).

Table 4: Annotation of trip segments

Name	Begin	End	Distance
Driving	2022-12-01 07:15:31	2022-12-01 07:54:53	6394
Driving	2022-12-01 08:11:51	2022-12-01 08:17:34	2910
Walking	2022-12-01 10:53:15	2022-12-01 10:58:30	392
Walking	2022-12-01 11:23:28	2022-12-01 11:23:28	352
Driving	2022-12-01 16:36:23	2022-12-01 11:27:44	7384

3.3 RP data Collection and Pre-processing

The necessary data to be collected have been set at a 30-day bundle of trips from 100 users per case study. To collect them, selected participants who have expressed interest to participate at the second phase of the TANGENT Travel Behaviour Survey have been contacted by email by the TANGENT team. The collection of their emails has been achieved by adding a dedicated question to the travel behaviour questionnaire, where interested readers were prompted to share their email with TANGENT researchers (Figure 5). To reach a large audience, the invitation to participate in the data collection was extended to TANGENT partners and their contacts, with a focus on partners who live and work in Athens, Lisbon, Manchester and Rennes.

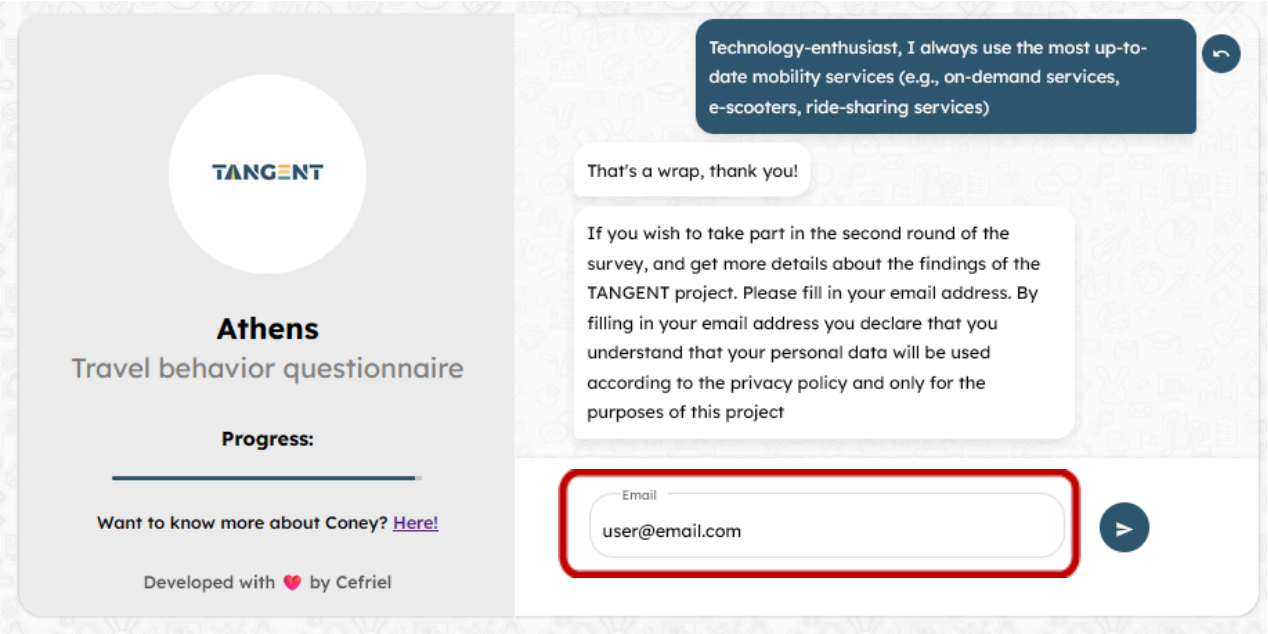


Figure 5: Collection of user emails through TANGENT's Travel Behaviour Survey

The collected emails were used to reach out to potential participants by sending them two leaflets. The first document contained detailed instructions on how to enable Google Timeline functionalities on a user's mobile phone across multiple mobile operating systems, how to download the data when the 30-day period has elapsed, and how to share their data with the TANGENT partners. The second document delineated participants privacy rights and TANGENT's data protection and confidentiality procedures. Participants have been also offered the opportunity to report complaints or grievances, or receive additional information on how their data are being treated. For this purpose, the email coordinates of NTUA's Data Protection Officer as well as those of the project coordinator have been shared with participants. A link to the Google data privacy & terms form has also been shared. For further reference, both documents are included in the Annex. Finally, a short video containing step-by-step visual and audio instructions on how to download the collected trajectories from a web browser environment was shared with potential respondents.

Data collection lasted for 10 months, from February 2023 to December 2023. The data collection progress and the overall dissemination activities have been actively monitored and coordinated across all four case study cities of TANGENT by both WP3 partners and case study leaders. To ensure a successful process, the WP3 team has regularly provided case study leaders with an updated overview of the collected sample and exchanged of past and planned dissemination activities. Additional to the initial contact, follow-up emails have been sent to remind potential users to activate Google Maps Timeline functionalities and/or to share their trajectories data, while TANGENT contact people have regularly been in direct contact with several of the participants, answering queries on the various steps of the data collection process. An overview of the dissemination activities is provided in Table 5.

Table 5: RP survey dissemination overview

Required Sample	30-days of commuters' trips from 100 commuters
Cities involved	All four case studies (Athens, Lisbon, Manchester, Rennes)

Survey duration	10 months
Start date	February 2023
Dissemination methods	by email; upon the collection of the data, participants are asked to forward them to the TANGENT team.

Figure 6 depicts the number of questionnaires and RP trips collected in each case study city.

Up until M28, data from a total of 991 trips made by 28 users were gathered, with the distribution across each case study city being as follows: Athens saw 729 trips from 22 commuters; Lisbon saw 120 trips from 3 commuters; Manchester contributed 142 trips from 3 users. No trips were collected in Rennes.

For the SP survey, a total of 2078 valid (excluding incomplete and fault answers) responses were collected for the four pilot cities through the questionnaire survey prepared within the framework of WP3. Specifically, 553 valid responses were collected for the case of Athens, 632 for Lisbon, 427 for Manchester and 466 for Rennes. A more detailed analysis of the sociodemographic characteristics of the sample is included in D3.2

In the pre-processing stage, in addition to the primary trip details derived directly from Google Maps Timeline data, we have further enhanced our analysis by calculating supplementary attributes for each journey, including its duration and associated costs.

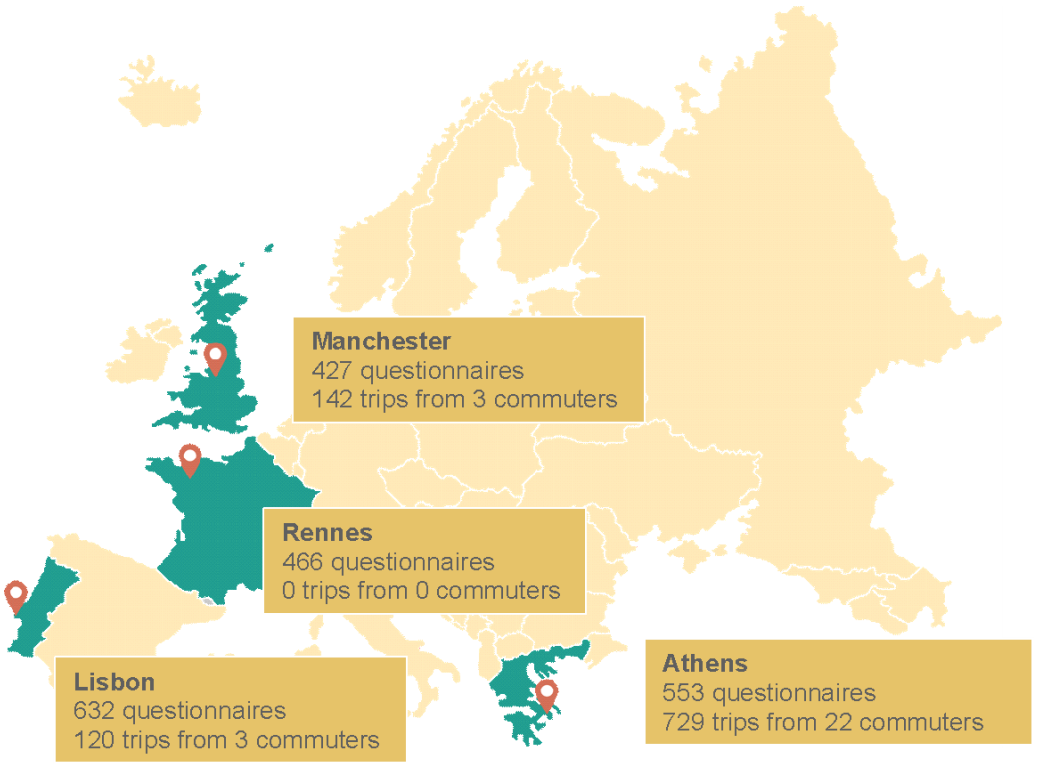


Figure 6: Collected samples per city

Trip duration was calculated by comparing the trip beginning and end timestamps. To calculate the **cost of trip**, and in the absence of any additional information on the specific characteristics of each participant's private car, it was assumed that every car consumes standard unleaded petrol at an average consumption rate of 10 litres per 100 kilometres. According to the latest available cost information at the time, a litre of unleaded petrol costs 1.86 € in Athens, 1.80 € in Lisbon, 1.75 € Manchester and 1.85 € in Rennes[1].

After this procedure, each RP trip for a particular trip is annotated by the following attributes:

- duration,
- distance,
- cost,
- origin and destination, and
- selected travel mode

Trip data were equally cleaned to eliminate erroneous or irrelevant entries. In the first category, the primary source of errors has been inaccurate tracking points, mainly due to the GPS signal failing to provide positioning information for a substantial part of a trip. One way to address this issue is by juxtaposing the activity staypoints (Table 3), to the trips between them (Table 4). If a journey doesn't start or end at a staypoint, it typically suggests that the GPS signal was lost at the beginning or end of the trip. Likewise, loss of signal during the trip is indicated when the GPS transmits sparse points with significant time intervals between them. This can be visually identified on a map as extended straight lines, which, when superimposed on a road network, are clearly unrealistic routes resembling lengthy "jumps". This problem has been resolved by establishing a minimum frequency (every 1 minute) for the GPS signal transmissions from the mobile device for each trip.

Non-urban trips were considered as irrelevant for the analysis and were thus filtered out. This is because the choices of transportation modes for these trips typically differ from those made within the city, due to varying availability of different transport options and other factors. Additionally, the longer distances and, subsequently, the duration of non-urban trips in comparison to the urban ones can act as outliers, affecting the overall fit of the developed models to the data. Finally, only the trips conducted to the city where participants live and work were retained, excluding urban trip to other cities that the individuals might have visited. A visualisation of a participant's relevant trips across the whole 30-day period is depicted in Figure 7.

[1] Prices for all cities retrieved from www.fuelo.net

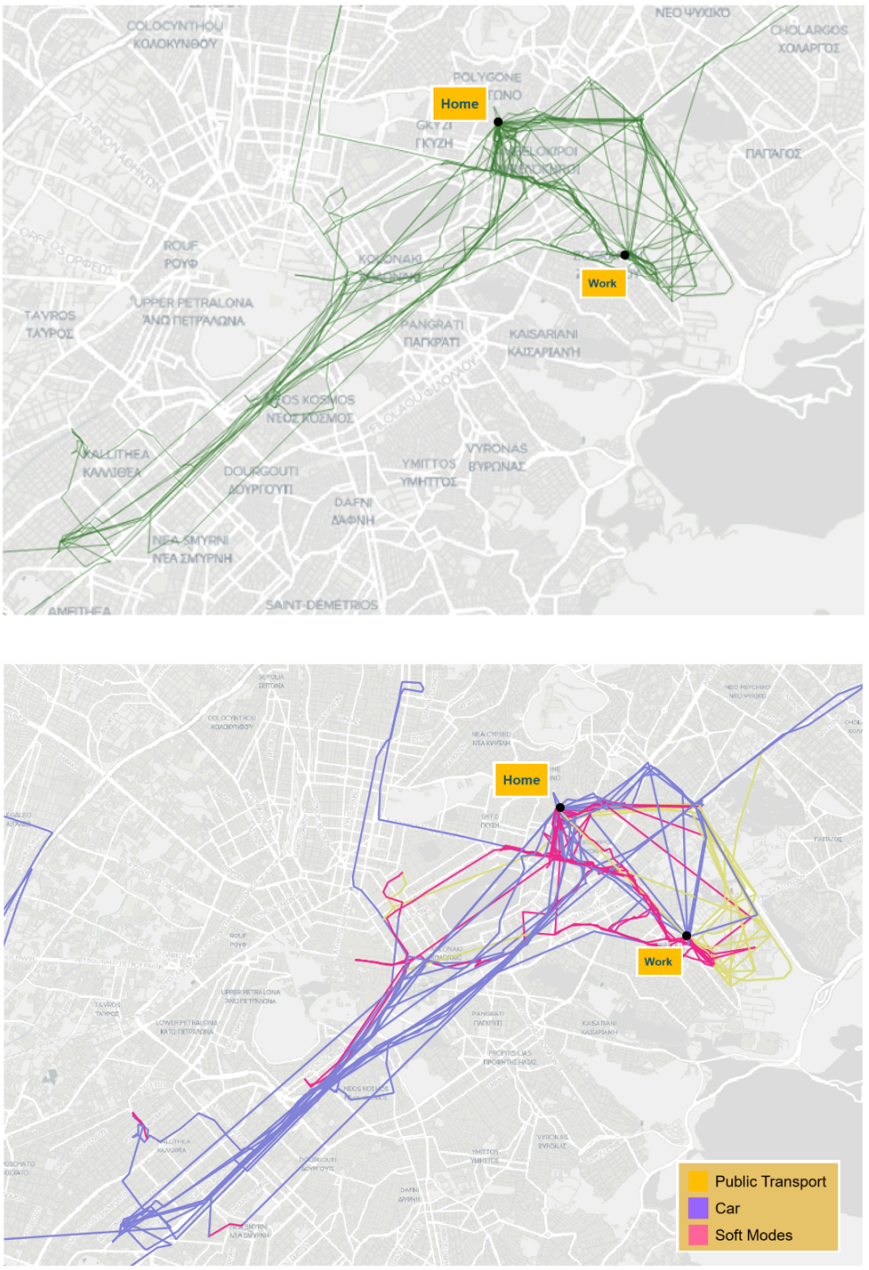


Figure 7: Visual representation of 30 days of trips for a commuter in Athens (left) and visual representation of travel mode choice for 30 days of trips for a commuter in Athens (right).

Broken down by mode, the same visualisation can be elaborated on the left side of Figure 7, where trips where public transport trips are shown in yellow, car trips in violet and soft mode trips (i.e. walking, riding a bicycle etc.) are shown in pink. Upon visual inspection, interesting mode choice patterns emerge regarding the spatial distribution of trips with each travel mode. In the example of Figure 7, selecting public transport or soft modes for home-work trips is preferred, comparing to more distant urban trips where a private car is mainly used. Of course, these patterns emerge only for a selected commuter, and are probably influenced by personal tastes, access to a private car, the

particularities of the home-work trip (both locations seem to be close enough to make on foot access feasible) and other, unobserved factors.

To gain insights on the overall behaviour of behaviour of the sampled commuters who have provided trip data, we proceed to an aggregate analysis of their trip characteristics, and the results are reported in Table 6.

Table 6: Aggregate characteristics of RP trips per transit mode

Transit Mode	Trip Count	Trip Count (%)	Average Trip Distance
Public Transport	230	14%	10.1 km
Private Car	754	47%	9.8 km
Soft Modes	611	38%	2.1 km

It is observed that private car trips account almost for half (47%) of the collected trips, followed by soft modes trips and public transport trips at 38% and 14% respectively. Regarding the average distance for each mode's trips, car and public transport trips are roughly equivalent in length, at 10.1km and 9.8 km. Interestingly, soft mode trips show an average distance of 2.1km, which appears to be somewhat inflated. We believe that this figure can be attributed to the particularities of the data collection method: it has been observed that Google Maps Timeline struggles to capture very short trips on foot, since their tracks fall in the vicinity of their origin staypoint, and as such, no movement is recorded.

3.4 Creation of the RP choices database

Following the method that has been outlined in the preceding sections, the database that contains the observed trip trajectories, for multiple users is created. Each trajectory is characterised by its origin and destination coordinates, the transit mode selected by the user and its characteristics with respect to travel time, cost etc. In other words, the database reveals the actual, observed, mode choice between various origin and destination pairs.

Even though we only observe the characteristics of a single mode option, we make the reasonable assumption that the commuter must have assessed the characteristics of all possible transportation modes before choosing the one we've recorded. To illustrate, when a commuter decides to travel from point A to point B using a private car, it implies that she has considered and subsequently rejected the public transportation options available in her area. To do so, she has assessed its attributes, such as travel time and cost, and found them less appealing than the option of using her private vehicle. Therefore, she chooses to use her private car for the journey. Although this mental procedure cannot be directly accessed as in the case of the highly structured stated preference scenarios, we make the case that it can be reasonably recreated, if accurate information on the non-chosen alternatives becomes available.

Conveniently, given an origin and destination pair, recent trip planner tools can provide us with very accurate estimations of the characteristics (travel time) of the non-chosen transit mode options. Consequently, our interest lies in recreating the synthetic choice scenarios that capture the decision-making process of commuters on the selection of the preferred transit mode in the case of

actual trips. In this work, we consider that each commuter has two options, either commuting by public transit, or by private car. This procedure can be conceived as follows:

- From a given trajectory, its origin, destination and the selected (actual) transit mode and its characteristics in terms of travel time and cost are derived.
- Given the origin and destination we are interested in recreating the characteristics of the non-chosen alternative. Intuitively, if it is observed that the trip was conducted by private car, we are interested in deriving the characteristics of the same trip conducted by public transit. Accordingly, if it is observed that the trip is done using public transport, the non-chosen alternative is the private car.
- To derive the characteristics of the non-chosen alternatives, a route planning tool is used.

A route planning tool is a digital application designed to help users find the most efficient path from one location to another. These tools typically use mapping data and the corresponding algorithms to offer directions for various modes of transportation. After inputting a starting point and a destination, the tools typically provide step-by-step directions, estimated travel time, and distance, while more advanced application also take into consideration real-time or past traffic information for the estimation of trip characteristics. In this work, we have used the Directions API[2] offered by the Google Maps platform. Given the coordinates of an origin and destination pair, the API calculates and returns the routes and estimated travel times for multiple transportation modes (driving, transit, cycling and walking), out of which we retain information for the first two.

At the end of this process, we have complete trip information for both the observed and the alternative mode choice and we are thus able to recreate the synthetic choice scenarios that capture a commuter's preference. Intuitively, the observed option is always chosen over the alternative option created using the route planner. Schematically, the process is depicted in Figure 8.

[2] More information for the Directions API is available at : https://developers.google.com/maps/documentation/directions

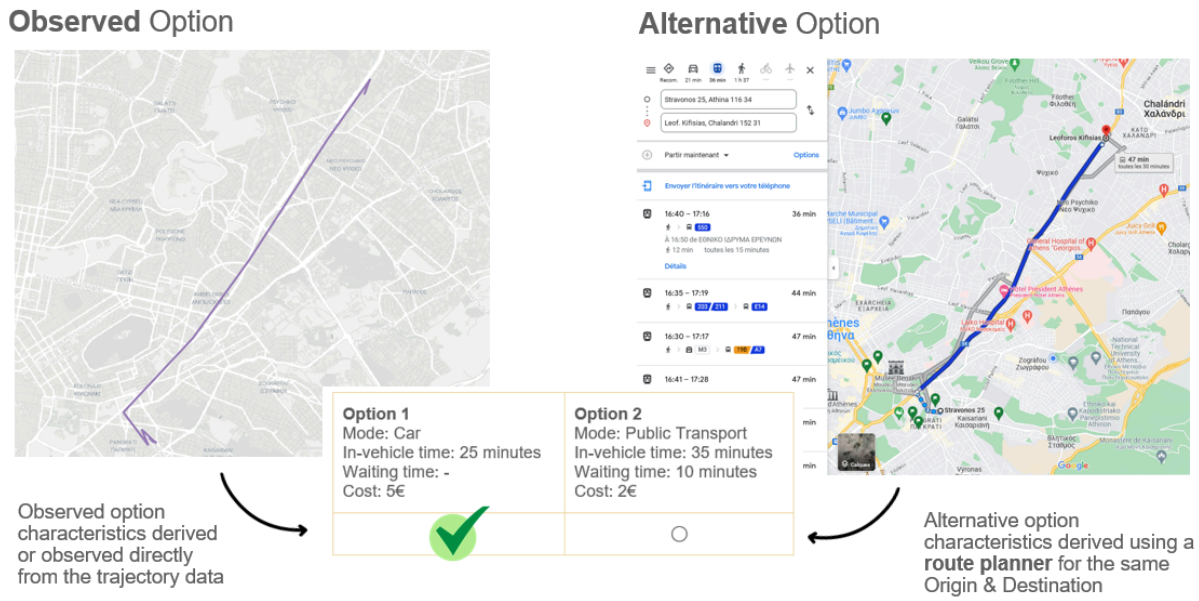


Figure 8: Creation of synthetic mode choices based on trajectory data

3.5 Creation of the joint SP-PR database

After the creation of the synthetic mode choice scenarios based on user revealed preferences, we end up with two different data sources.

On the one hand, we have a database of commuters' stated mode preferences when novel traffic management strategies are applied in an urban setting in the future.

On the other hand, the procedure that was described in the previous section provides us with a series of revealed preferences scenarios that elicit actual mode choices in a real-world setting where such strategies have not been implemented yet.

Our interest lies in merging these two databases, integrating insights from both hypothetical scenarios (stated preferences) and real-world choices (revealed preferences) to get a more comprehensive understanding of commuter behaviour. As discussed in Chapter 2, this approach combines the capacity of hypothetical scenarios to elicit behaviours that cannot be observed in the market, with the practical evidence of actual choices that serve as an "anchoring to reality" for the stated preferences.

To merge the two databases, we need to address the following issues.

The first issue is related to the fact that, in reality, each of the databases refers to different network conditions, since in the case of the SP scenarios traffic management strategies are applied and render some modes more or less attractive for commuters. In the case of the RP scenarios, mode choices are not affected by such strategies since they have not been implemented yet, but, interestingly, they can be affected by other strategies or factors. We make the case that the choices that commuters take are, in fact, *strategy-agnostic*. That is, commuters only evaluate the

characteristics of each mode with respect to, for example, commute time and cost before opting for either a private car or the public transport. In other words, trade-off ratios among attributes remain the same, regardless of the setting (SP or RP) where this decision is taken.

This allows for the "stacking" of both SP and RP database and the joint use of both for the creation of a model that are trained and tested on both data types simultaneously.

Be that as it may, the two SP & RP databases are, naturally, derived from the quite different data generation processes that have been described in Section 3.1.2. The literature on data fusion acknowledges the need of aligning the measurement attributes of various elicitation methods when integrating data sources. In particular, it is theoretically essential to adjust for differences in error variance of the different dataset when merging multiple choice data sources[3]. This consideration is addressed by estimating a scaling factor that accounts the relative error variance of one source with respect the other.

This operation is part of the model calibration and will be described in detail in Section 4.1. For now, we are interested in retaining the information of whether an observation or data point comes from the RP or the SP dataset. This can be achieved by adding a dummy binary variable indicating whether the variable comes from the RP database (taking the value 1), or from the SP database (taking the value of 0).

Finally, we note a certain imbalance of information between SP and RP data points. While in RP data points we only have information on the preferred mode across a series of choice scenarios, the information derived from the stated behaviour questionnaire is richer, including questions that elicit the mobility profile, commuters' perception on the various traffic management strategies of TANGENT, and sociodemographic characteristics of respondents.

When merging the two databases, we are interested in enriching each RP choice situation with information about the commuter conducting it, in order to add context to the choices individuals make and assure that each datapoint of the joint SP-RP database is characterised by the same explanatory variables.

In this work, due to the procedure that we have used to collect the RP preferences, it has become possible to associate each RP choice situation to a specific user and then to her stated behaviour questionnaire. This has been achieved by leveraging the common information inside both the questionnaire and the Google Maps Timeline files. For instance, several survey participants have inputted their email at the end of the stated preference questionnaire, while each point and trajectory downloaded by the Google Maps Timeline also contains in its metadata the email that each user uses to connect to Google services. Using this common information, we have been able to enrich each RP choice situation with information about the commuter conducting it. This information can be used to improve the accuracy of the developed travel choice models and assure that RP and SP data points have a similar structure.

[3] Morikawa T. Incorporating stated preference data in travel demand analysis..Ph.D. Dissertation, Department of Civil Engineering, Massachusetts Institute of Technology, 1989.

4. Model Development and Evaluation

4.1 Travel choice model specifications

We proceed in creating travel mode choice models that are trained and tested on a combined Stated Preference (SP) and Revealed Preference (RP) database, with a twofold goal: firstly, to accurately forecast the preferred mode of travel when TANGENT's traffic management strategies are implemented, and secondly, to effectively incorporate and manage data derived from various sources. Given that very few data was collected for each of Lisbon, Manchester and Rennes, the models have been created based on the joint SP-RP database for Athens, for both “Prioritisation” and “Pricing” strategies.

In this study, parametric Multinomial Logistic Regression Models have been selected as the optimal modelling choice, primarily because they can accommodate robustly both Stated Preference (SP) and Revealed Preference (RP) data in their specifications. Multinomial logit models are a type of statistical model used to predict the probability of different outcomes in situations where there are multiple, mutually exclusive choices. They are an extension of the binary logistic regression model and are particularly useful in analysing categorical data where the dependent variable can take on more than two categories. At the heart of the multinomial logit model is the concept of utility. Each alternative is assumed to be associated with a utility for the individual, and the choice made by them is the one from which she derives the highest utility. The exact utility of each alternative is not directly observable but it is modelled as a linear combination of observed variables and a random component. The observed variables can be characteristics of the alternatives (such travel time and cost), characteristics of the individual (such as the information capture in Sections A, B and D of the TBS) or both. The estimated coefficients of these variables represent the influence of each factor on the preference for an alternative.

To mix data from different sources, a scale factor is needed to equate the random error variances associated with each data type (Louviere et al., 2000). The intuition behind this is that when RP and SP data are combined and estimated jointly, it is typical for the error terms of the different data sets to exhibit unequal variances. Typically, SP data have a higher variance, attributed sometimes to the difficulty of expressing consistent preferences in hypothetical scenarios where there is no hands-on experience.

In this situation, while one data set's variance is normalized to one, the relative variance (or 'scale') of the other data set is identified and can be estimated. In the majority of applications, the RP data are considered to be of the “correct” scale, since they are associated with behaviours identified in the “real market” (Brownstone, 2000). Then an “SP scale” coefficient is multiplied with all SP data to normalize the variances in the random components of the utility functions. This “SP scale” is hereafter denoted as σ. The interpretation of the scale parameter will further be further discussed in Chapter 4, where the joint estimation outputs are further discussed (pages 31 and 33).

Taking into account these considerations, the utility obtained by individual i from selecting alternative j in choice situation t in either a hypothetical (SP) or real-life (RP) scenario, is defined as equation:

$$u_{ijt} = \beta' x_{ijt} + \gamma' y_{ijt} + \frac{\varepsilon_{ijt}}{\sigma}, \#(1)$$

where u_{ijt} is the utility commuter i receives from selecting mode j, $\beta^{'}x_{ijt} + \gamma^{'}y_{ijt}$ is the systematic portion of the utility, $\beta^{'}$ and $\gamma^{'}$ are vectors of parameters, x_{ijt} is the vector of variables specific to the alternative characteristics, y_{ijt} in the vector of variables specific to the individual characteristics, ε_{ij} is a random variable that accounts for the effects on preferences of unobserved attributes of the alternative and individual. This specification allows the model to capture how individual characteristics interact with the attributes of each choice to affect the overall utility. Additionally, σ is the standard deviation of the error term, i.e., the scale parameter. Note that, if the choice situation t happens in an RP setting, the scaling coefficient is set to 1. Otherwise, it is a parameter to be estimated.

Multiplying by it, the utility for an individual commuter becomes:

$$u^{'}_{ijt} = \sigma\beta^{'}x_{ijt} + \sigma\gamma^{'}y_{ijt} + \varepsilon_{ijt} \#(2)$$

The multinomial logit (MNL) model assumes the error terms in Equation 2 are independent and identically distributed (iid) extreme value Gumbel variates; consequently, Logit choice probability that an individual i selects mode j in choice situation t is represented as follows:

$$\pi_{ijt} = \frac{\exp exp\left(\sigma\beta^{'}x_{ijt} + \sigma\gamma^{'}y_{ijt}\right)}{\sum_{j=1}^{J} \exp exp\left(\sigma\beta^{'}x_{ijt} + \sigma\gamma^{'}y_{ijt}\right)} \#(3)$$

In terms of estimation, the coefficients in a multinomial logit model are usually estimated using maximum likelihood estimation. This involves finding the set of coefficients that make the observed choices most probable under the model. To jointly estimate the scaling coefficient σ, the SP data are iteratively rescaled until the joint likelihood is maximized (Brownstone & Train, 1998)

4.1.1 Model evaluation

Model evaluation involves applying various metrics to assess the performance of an econometric (or other) model on its capacity to perform the designed task. Good performance across a range of metrics is a robust indicator of the model's reliability and suitability as a decision-making tool, while they also serve as a validation of the model's specifications.

Typically, the evaluation pipeline of a classification model involves training a model on a part of the dataset, applying it on the rest of the dataset and assessing its classification performance through the estimation of metrics. By testing the model on new data, the model's ability to generalize on unseen data is assessed, while overfitting, or the exact fit of the model on the used training data, is avoided.

Quality assessment metrics for classification are derived from a confusion matrix, which tabulates both accurately and inaccurately identified instances for each category, illustrating the algorithm's effectiveness in distinguishing between accurate and inaccurate classifications. The simplest version of a confusion matrix for a binary classification activity is depicted in Table 7, and can be easily generalised for classification models with multiple categories.

Table 7: A confusion matrix for binary classification

Actual Class / Predicted Class	As Positive	As Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

In binary classification, there are two possible outcomes: positive and negative. Consider a scenario where a classification model is used to predict an individual's mode of choice between car and public transit. In that case, the positive class could be arbitrarily set at choosing public transit, with the negative outcome being choosing a private car. Each cell of the confusion matrix visualises the performance of the algorithm in classifying these outcomes.

- **True Positives (TP)**: Instances where the model correctly predicts that an individual would choose public transit.
- **False Positives (FP)**: Instances where the model incorrectly predicts that an individual would choose public transit when they actually choose a car.
- **False Negatives (FN)**: Instances where the model incorrectly predicts that an individual would choose a car when they actually choose public transit.
- **True Negatives (TN)**: Instances where the model correctly predicts that an individual would choose a car.

Measures of the classification performance are derived directly from the confusion matrix. The most used empirical measure, accuracy, does not distinguish between the number of correct labels of different classes, and is defined as:

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \#(4)$$

Accuracy can be interpreted as a metric for how often the model is correct in its prediction. Focusing on one class, the metrics of $Precision$, $Recall$, and $F1 - score$ can be estimated. Among a group of classes, one particular class is identified as the primary focus (typically labelled as positive). The remaining classes are treated in one of two ways: either they are maintained as distinct entities for multi-class classification, or they are combined into a single class for binary classification. The preferred metrics are computed for the positive class, although the same measures can be computed for the negative class.

$$Precision = \frac{TP}{TP+FP} \#(5)$$

$$Recall = \frac{TP}{TP+FN} \#(6)$$

$$F1 - score = \frac{2\times Precision \times Recall}{Precision + Recall} \#(7)$$

Returning to the example where public transit is chosen as the positive class, the $Precision$ metric can be interpreted as a measure of how many of the predicted public transit trips were actually done with public transit. $Recall$ is interpreted as a measure of how many of the actual public transit trips were

predicted correctly. Combining the two, the $F1 - score$ is the harmonic mean of $Precision$ and $Recall$ for the positive class. In the multi-class and multi-label case, each of the metrics is calculated for each of the classes and can be averaged into an overall indicator.

4.2 Estimation Results

Applying the methodology of Chapter 3, the parameters of two Multinominal Logit Models, for each of the "Prioritisation" and "Pricing" strategies have been estimated using the Maximum Likelihood Estimation (MLE) approach.

4.2.1 Estimation results for "Prioritisation"

For the "Prioritisation" strategy, Table 8 below summarizes the explanatory variables that have been used to model mode choice decisions for a trip to the city centre, where the two available options are either the private car, or public transit.

Table 8: MNL model variables for the "Prioritisation"

Input Variables	Description	Type	Unit
IN-VEHICLE TIME	In-vehicle time	Integer	min
WAIT TIME	Waiting time	Integer	min
COST	Trip cost	Integer	Euro
FREELANCER	The respondent self-reports his income as high	Binary [Yes-No]	-
RETIRED	The respondent is retired	Binary [Yes-No]	
LOW INCOME	The respondent self-reports his income as low	Binary [Yes-No]	-
PASSENGER SATISFACTION	Satisfaction with the level of service of Public Transport	5-point scale [Not satisfied at all - Very satisfied]	-
HOUSE MEMBERS	Number of members in the individual's household	Integer	Individuals

Three alternative-specific variables are considered: In-vehicle time, or the time spent inside a vehicle during a trip, waiting time or the duration that passengers spend waiting before they are transported and the trip cost. The waiting time for the use of a private car is presumed to be zero and the total travel time for each mode is composed by both the in-vehicle time and the waiting time. Alternative-specific variables vary not only across individuals but also across the different alternatives (modes of transport) available to each individual. They also coincide with the variables contained in each of the choice situations assessed by participants for this strategy (see Table 1).

In addition to the alternative-specific variables, individual-specific variables are unique to each individual but are constant across the different transportation alternatives. The model contains five such variables. Two of the variables are related to an individual's occupation, transcribing whether she works as a freelancer, or she is retired. A third variable indicates if an individual's income falls into the "low" category, meaning her annual earnings do not exceed 10,000 euros. Passenger Satisfaction expresses the level of contentment or approval expressed by passengers regarding their travel experience while using the PT system on a 5-point scale. The final variable (House members) captures the number of individuals in each respondent's household.

Table 9 summarizes the estimation results for the MNL model developed for “Prioritisation”. Each coefficient represents the change in the log odds of the dependent variable being in a particular category (compared to the reference category) for a one-unit increase in the predictor variable, holding all other variables constant. In this case, the reverence category is the selection of public transport, meaning that a positive coefficient expresses an increase in the likelihood of selecting a private car over selecting public transport.

Table 9: MNL estimation results for “Synchronisation”

Feature	Coefficient	P-value
Intercept PT	0.199	0.522
IN-VEHICLE TIME	-1.169	0.000
WAIT TIME	-6.271	0.000
COST	-0.212	0.000
FREELANCER	0.453	0.007
RETIRED	0.375	0.000
LOW INCOME	-0.305	0.034
PASSENGER SATISFACTION	-0.155	0.000
HOUSEHOLD MEMBERS	0.100	0.000
SP Scale	1.676	0.000

It is observed that the direction of the effect of all coefficients is the expected. Increases in the in-vehicle and wait time, and travel cost lead to a decrease in the probability of selecting a mode over the alternative. Regarding the effects of individual-specific characteristics, low income is intuitively correlated with a lower likelihood of choosing one's car over public transportation. Additionally, the more satisfied someone is with the public transportation network, the less likely they are to choose their car in relation to public transportation. Conversely, the more members there are in a household, the greater the likelihood of using a car. The same applies to individuals who work as freelancers or are retirees. Finally, the value of the intercept suggests that, all things equal, public transit is preferred over the private car.

In the joint estimation of Stated Preference (SP) and Revealed Preference (RP) data, the scale parameter that expresses the difference in variance between the random error terms of the two data types has been estimated at 1.676, suggesting that the variance due to unobserved factors in the SP data's utility is 67.6% greater than that in the RP data. It is equally statistically significant, indicating that this difference does indeed exist. The fact that that SP data tend to have more variance than RP data also coincides well with the findings of other works that have conducted joint RP/SC analyses (Börjesson, 2008).

The p-values of the coefficients provide a measure of the statistical significance of each variable. The p-value of each coefficient represents the probability of obtaining the observed results, assuming that the null hypothesis that that the observed coefficient has no effect or no difference is true. A lower p-value indicates that the observed data are less likely to occur under the null hypothesis and if the p-value is less than the chosen significance threshold of 5% or 0.05, the variable is considered as

statistically significant. At the exception of the intercept, all explanatory variables selected for inclusion in the final model are statistically significant.

The developed model was finally evaluated by estimating the $Accuracy$, $Precision$, $Recall$ and $F1-score$ metrics, reported in Table 10. From our experience from models developed for similar mode choice problems, these results are satisfactory.

Table 10: MNL model quality metrics for "Prioritisation"

Metric	Value
Accuracy	0.68
Precision	0.69
Recall	0.67
F1-Score	0.67

4.2.2 Estimation results for "Pricing"

The equivalent analysis is also conducted in the case of "Pricing". In this case, there are three available mode choice options to access the city centre: select the private car and pay the toll price, select the private car and bypass the area where the congestion pricing scheme is applied in order to avoid paying the toll or, alternatively, use the available public transit options for the trip. Table 11 reports the explanatory variables that were included in its specifications.

Table 11: MNL model variables for the "Prioritisation"

Input Variables	Description	Type	Unit
TIME	Total travel time	Integer	min
COST	Total travel cost (including the toll, where it applies)	Integer	Euros
UNEXPECTED	Respondent's self-evaluation of the probability to change the usual travel mode in case of hazardous events (e.g., COVID-19)	5-point scale [Not likely at all - Certain]	
CITY CENTER	The respondent lives in the city centre	Integer	
HIGH INCOME	The respondent self-reports his income as high	Binary [Yes-No]	
GENDER	The respondent's gender	Binary [Male-Female]	
FREELANCER	The respondent is employed as a freelancer.	Binary [Yes-No]	

Two alternative-specific variables are considered in this case: the total travel time for a given trip and its cost. The rest of the variables are individual-specific and capture each individual's response to unexpected hazardous events, whether she lives in the city centre of Athens, as well as her income, gender, and whether she is employed as a freelancer. Similarly, the "Prioritisation" scenario, an individual is considered to have a high income if it surpasses 25,000 euros in Athens.

The estimation results of the MNL model for pricing are reported in Table 12. In this case, the option of selecting public transport is designated as the reference group, affected only by the two alternative-specific (trip time and cost). The coefficients for the two other outcome groups (selection of car with and without toll payment) are estimated according to this class.

Table 12: MNL estimation results for "Pricing"

Independent Variable	CAR w/ TOLL		CAR w/o TOLL	
	Coefficient	P-value	Coefficient	P-value
TIME	-2.614	0.000	-2.614	0.000
COST	-0.255	0.000	-0.255	-0.255
Intercept	0.246	0.157	0.408	0.000
UNEXPECTED	-0.096	0.017	-0.102	0.004
CITY CENTER	-0.373	0.000	-0.466	0.000
HIGH INCOME	0.769	0.000	0.122	0.035
GENDER	0.211	0.033	0.191	0.027
FREELANCER	0.793	0.000	0.598	0.000

It is observed that an increase of travel time and cost of an alternative affects negatively the likelihood of choosing this alternative compared to the others. Regarding the individual-specific variables that codify the characteristics of each respondent, identifying as male also increases the probability of selecting a car over public transit, whether or not a toll is involved, by a similar margin. The result is similar for individuals employed as freelancers, as well as for those earning a high income. Interestingly though, it can be observed that for high-income individuals, the magnitude of the change in the log odds of selecting the car and paying the toll over selecting PT (0.769) compared to the equivalent value for the option of not paying the toll (0.122) over PR is higher. This aligns with the typical belief that individuals with higher incomes place greater value on their time and are therefore more likely to pay for a toll. In contrast, individuals who live in the city centre are less likely to choose a car over public transport, whether they pay the toll or not, as there are individuals who are more inclined to change their usual mode of transport in the case of an unexpected event. All coefficients are statistically significant at the 0.05 level. The scale factor was estimated at 1.131 and is statistically significant, indicating that there is indeed a difference in variance between the random error terms of the SP and RP observations, with SP data displaying more variance that RP data.

Finally, the model underwent an evaluation process to assess the accuracy of its predictions. The dataset was divided, with 80% allocated for training and the remaining 20% used for testing. The evaluation metrics – $Accuracy$, $Precision$, $Recall$, and the $F1-score$ – were computed and their results are presented in Table 13.

Table 13: MNL model quality metrics for "Pricing"

Metric	Value
Accuracy	0.45
Precision	0.48
Recall	0.44
F1-Score	0.43

These scores, while relatively low, are on par with the quality metrics achieved by the MNL model trained solely on SP data.

5. Conclusion

The present study aimed to investigate the impact new modes and traffic management strategies may have on mode attractivity and on the subsequent mode choices of commuters. The main advantage of the present work is the exploitation of two complementary data sources: stated preference data collected using a questionnaire survey and revealed preference data, obtained by analysing actual commuter choices.

For the development of mode choice models, two distinct scenarios were examined for each city: (1) choice between car and Public Transport when prioritization is given to public transport over car, and (2) a choice between car, car with toll payment, and public transport based on the cost of the trip.

SP data was collected by distributing TANGENT's travel behaviour survey. The cornerstone of the survey has been the stated behaviour scenarios, where participants are presented hypothetical scenarios and are requested to select the most attractive mode alternative. Each questionnaire also documents users' mobility profiles, perception of users towards new services and traffic management strategies, and sociodemographic characteristics.

The SP data collected through the travel behaviour survey were enriched with RP data based on actual mode choice decisions made in real situations. Trip trajectory data was collected from interested participants, who have exported and shared with the TANGENT team the travel data accumulated from the Google Maps app on their smartphones over a period of one month. By analysing the transport modes that commuters selected between various starting points and destinations, and by modelling the attributes of the alternatives they did not choose, the data were used to reconstruct the users' real travel decisions.

Both databases have been then combined in a joint SP-RP. This approach facilitates the exploitation of the advantages each data source offers, as well as the correction of inherent biases present in each dataset. Further methodological issues on the fusion of different data sources and creation of joint SP-RP models were extensively discussed.

To identify the most prominent factors that influence mode choice, multinomial logit models were trained and tested on the joint databases. The specification of each model comprehensively accounts for the fact that the data come from different data sources, by the inclusion of a scale factor that adjusts variance differences between SP and RP data. After their creation, the models' predictive capabilities were assessed through the computation of accuracy, precision, recall, and F1-score metrics, and the results indicated strong performance for both models.

The modelling results indicate that increases in travel time and cost diminish the perceived utility of every mode. Moreover, factors like an individual's socioeconomic profile (occupation, income, gender, etc.), preferences, and perceptions of each mode's quality, their reaction to unexpected events, and others, were shown to significantly impact travel choices in a statistically meaningful way.

Some limitations of our work include the relatively few RP data, which could lead to a bias in the results towards the more abundant SP data. Increased data availability could provide a deeper insight into individuals' transportation choices and allow for the application of this methodology to the datasets from cities like Lisbon, Manchester, and Rennes, where currently the data is too limited for effective model calibration.

We also note that the MNL logit models that have been created assume that the Independence of Irrelevant Alternatives (IIA) assumption holds, or that the relative odds of preferring one option over

another should remain constant regardless of whether a third, irrelevant alternative is available. In practice though, this assumption can be invalidated in transit mode choice situations through a phenomenon often referred to as the "red bus/blue bus" problem. Suppose for instance that a new congestion pricing policy is introduced in a case where commuters originally have two choices: driving a car (without paying the congestion toll) or using public transit. According to the IIA assumption, the introduction of this new option (driving with toll) should not affect the relative preference between the original two options (driving without toll and public transit). Intuitively thought, the introduction of the toll-free driving option might disproportionately affect the choices of car users than public transit users.

Further research on this topic could evaluate other random utility models that handle better these considerations. A possible candidate are Mixed logit models that allow random coefficients in their specifications, offering a way around multinomial models' IIA assumption. A second model could be Nested Logit extension of the standard logit mode, which allows for the grouping of similar choice alternatives into 'nests', facilitating varying levels of similarity among groups of options within a set of choices. This feature is particularly useful when alternatives are not entirely distinct and share common characteristics. For instance, in the case of a congestion pricing scenario the two car-related options could be included in the same nest.

Closing, we note that accurate predictions of the modal split have significant potential to assist policymakers and transportation planners in promoting sustainable transportation modes. By estimating the impact of new services and traffic management strategies on travel demand for different modes, these predictions can help identify the factors that influence commuter behaviour and design effective strategies accordingly.

6. References

Ben-Akiva, M., Bradley, M., Morikawa, T., Benjamin, J., Novak, T., Oppewal, H., & Rao, V. (1994). Combining revealed and stated preferences data. Marketing Letters, 5(4), 335–349. https://doi.org/10.1007/BF00999209

Börjesson, M. (2008). Joint RP–SP data in a mixed logit analysis of trip timing decisions. Transportation Research Part E: Logistics and Transportation Review, 44(6), 1025–1038. https://doi.org/10.1016/j.tre.2007.11.001

Brownstone, D., & Train, K. (1998). Forecasting new product penetration with flexible substitution patterns. Journal of Econometrics, 89(1–2), 109–129. https://doi.org/10.1016/S0304-4076(98)00057-8

Hanley, N., & Czajkowski, M. (2019). The Role of Stated Preference Valuation Methods in Understanding Choices and Informing Policy. Review of Environmental Economics and Policy, 13(2), 248–266. https://doi.org/10.1093/reep/rez005

Huang, Y., Gao, L., Ni, A., & Liu, X. (2021). Analysis of travel mode choice and trip chain pattern relationships based on multi-day GPS data: A case study in Shanghai, China. Journal of Transport Geography, 93, 103070. https://doi.org/10.1016/j.jtrangeo.2021.103070

Ilahi, A., Belgiawan, P. F., Balac, M., & Axhausen, K. W. (2021). Understanding travel and mode choice with emerging modes; a pooled SP and RP model in Greater Jakarta, Indonesia. Transportation Research Part A: Policy and Practice, 150, 398–422. https://doi.org/10.1016/j.tra.2021.06.023

Lavasani, M., Hossan, M. S., Asgari, H., & Jin, X. (2017). Examining methodological issues on combined RP and SP data. Transportation Research Procedia, 25, 2330–2343. https://doi.org/10.1016/j.trpro.2017.05.218

Lizana, P., Ortúzar, J. D. D., Arellana, J., & Rizzi, L. I. (2021). Forecasting with a joint mode/time-of-day choice model based on combined RP and SC data. Transportation Research Part A: Policy and Practice, 150, 302–316. https://doi.org/10.1016/j.tra.2021.06.006

Louviere, J. J., Hensher, D. A., & Swait, J. D. (2000). Stated choice methods: Analysis and applications. Cambridge University Press.

Mark, T. L., & Swait, J. (2004). Using stated preference and revealed preference modeling to evaluate prescribing decisions. Health Economics, 13(6), 563–573. https://doi.org/10.1002/hec.845

Morikawa, T. (1994). Correcting state dependence and serial correlation in the RP/SP combined estimation method. Transportation, 21(2), 153–165. https://doi.org/10.1007/BF01098790

Small, K. A. (1987). A Discrete Choice Model for Ordered Alternatives. Econometrica, 55(2), 409. https://doi.org/10.2307/1913243

Tabasi, M., Raei, A., Hillel, T., Krueger, R., & Hossein Rashidi, T. (2023). Empowering revealed preference survey with a supplementary stated preference survey: Demonstration of willingness-to-pay estimation within a mode choice case. Travel Behaviour and Society, 33, 100632. https://doi.org/10.1016/j.tbs.2023.100632

Yan, X., Levine, J., & Zhao, X. (2019). Integrating ridesourcing services with public transit: An evaluation of traveler responses combining revealed and stated preference data. Transportation Research Part C: Emerging Technologies, 105, 683–696. https://doi.org/10.1016/j.trc.2018.07.029

Annex

TANGENT

Enabling & Sharing Google Timeline Data

Travel Behavior Survey

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 955273

What is the Google Maps Timeline?

Google Maps Timeline shows an estimate of places that you may have been and routes that you may have taken based on your Location History. You can find how far you've traveled and the way that you traveled, such as walking, biking, driving, or on public transport. You can learn more about Google Timeline by clicking this link.

Google Maps Timeline data is stored in your Google account and can be exported and shared. For the purposes of our research conducted within TANGENT, we are kindly asking you to **export and share** with the TANGENT team (surveys.tangent@gmail.com) **the trips of a typical month (30 days)** of your Google Maps Timeline**.** The duration of this process is estimated at **5 minutes.**

How to enable Google Maps Timeline functionalities

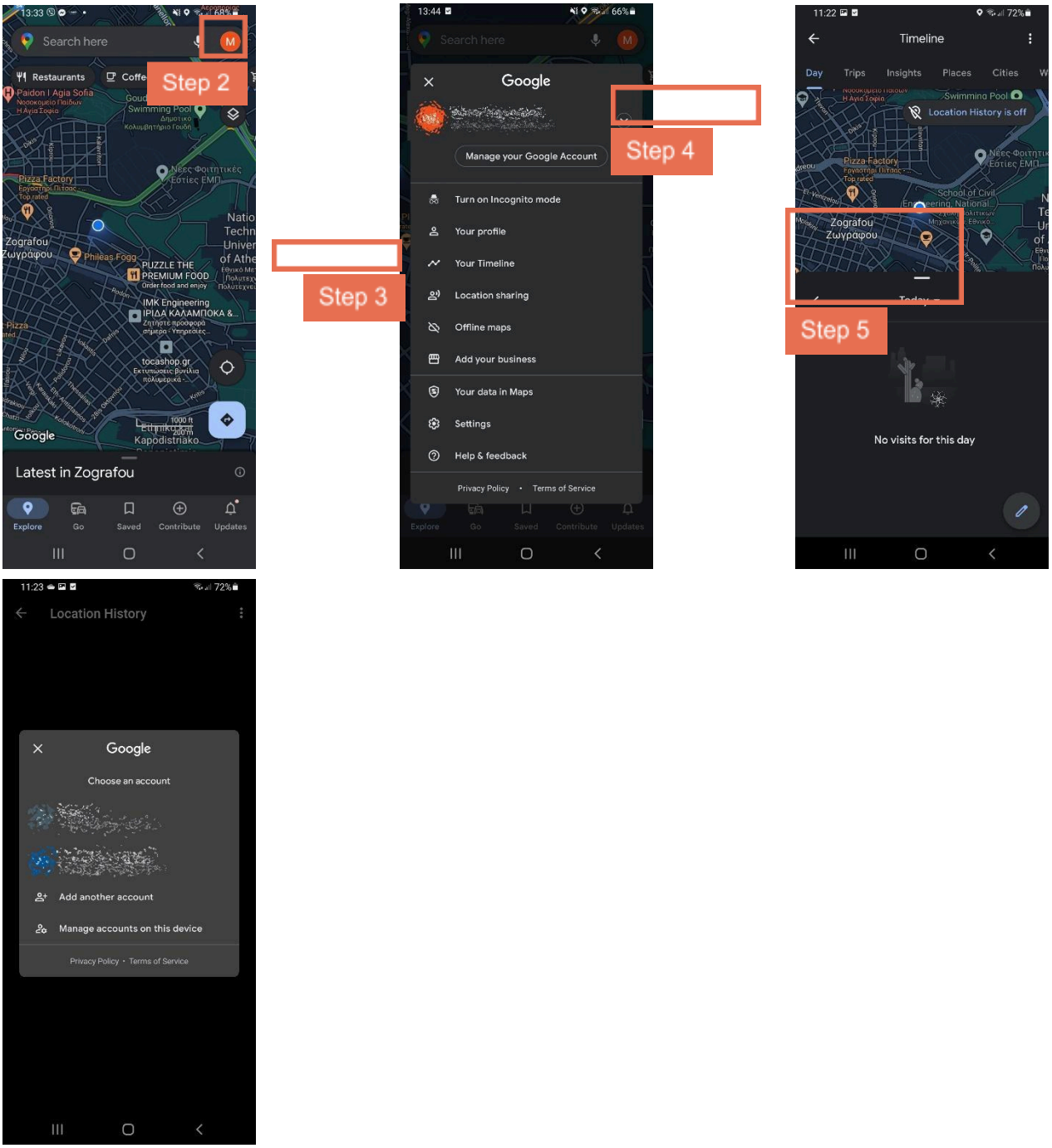
Follow the instructions that follow to ensure that Goole Maps Timeline functionalities are enabled and activate them if they are not.

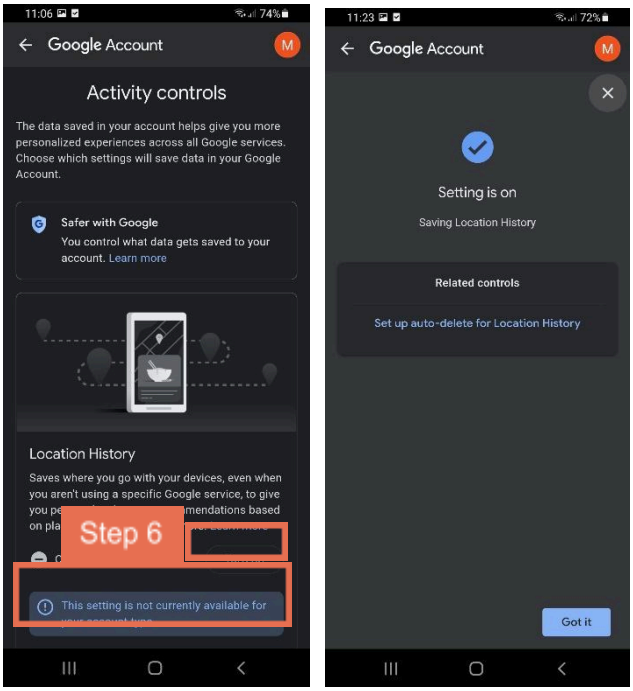
On your mobile phone

1. On your mobile phone, open the Google Maps App 
2. At the top right, tap your profile picture  or initial 
3. Tap: Your Timeline 
4. At the top right, tap: **Location History is off**
5. Choose your Account (if needed)
6. Tap: **Turn on**

In case the "Turn on" option is not available and the message "This setting is not currently available" is displayed below it, you may need to verify your age by following the instructions here.

Visual Guide:



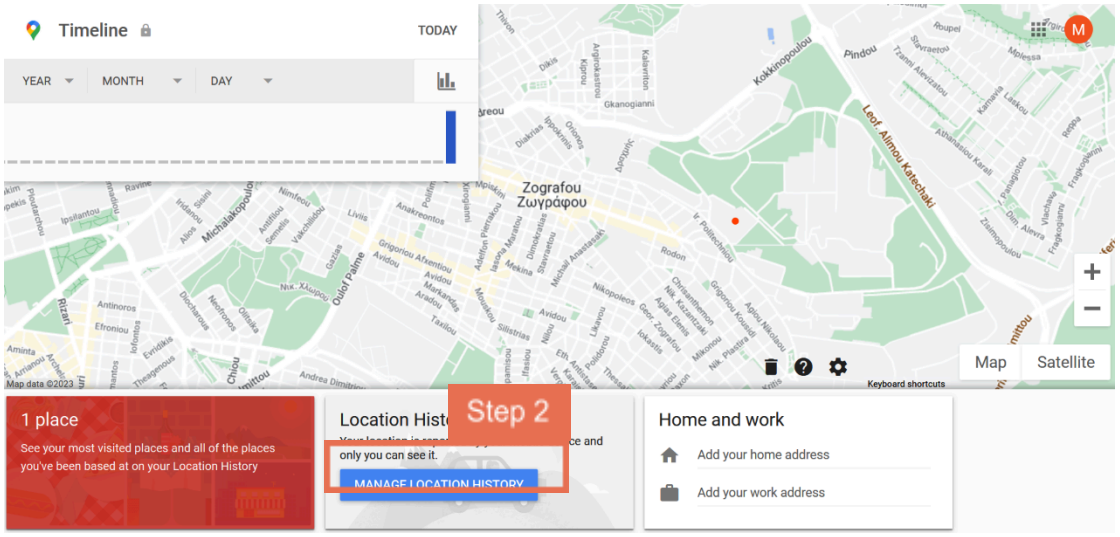

11:06
74%
Google Account
Activity controls
The data saved in your account helps give you more personalized experiences across all Google services. Choose which settings will save data in your Google Account.
Safer with Google
You control what data gets saved to your account. Learn more
Location History
Saves where you go with your devices, even when you aren't using a specific Google service, to give you pe mendations based on pla
Step 6
This setting is not currently available for
11:23
72%
Google Account
Setting is on
Saving Location History
Related controls
Set up auto-delete for Location History
Got it

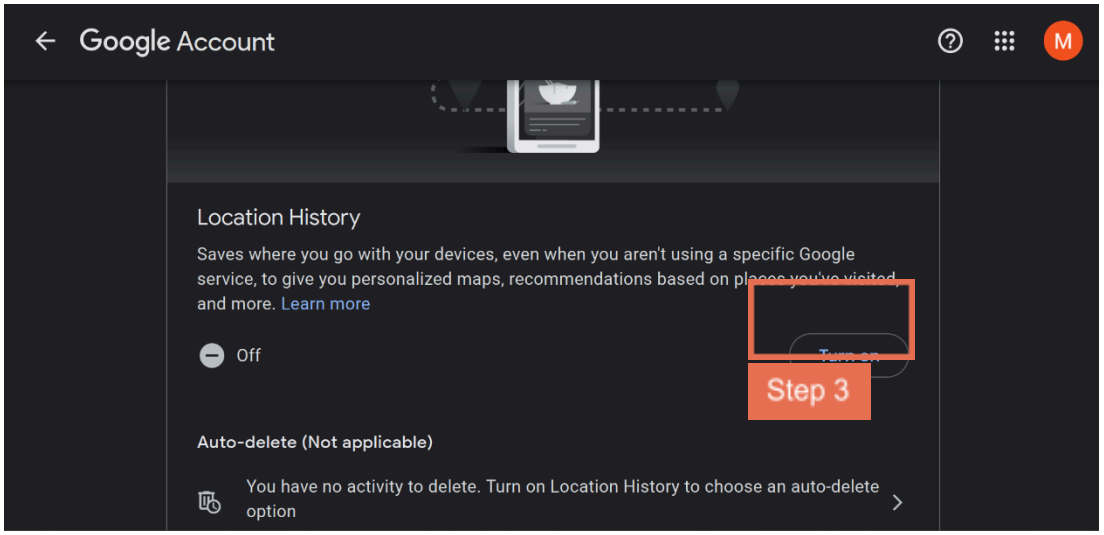
On your computer

1. On your computer, go to Timeline
2. At the bottom of the page, click on: Manage Location History
3. Scroll down to the “Location History Section” and click: Turn on

In case the "Turn on" option is not available and the message "This setting is not currently available" is displayed below it, you may need to verify your age by following the instructions here.

Visual Guide





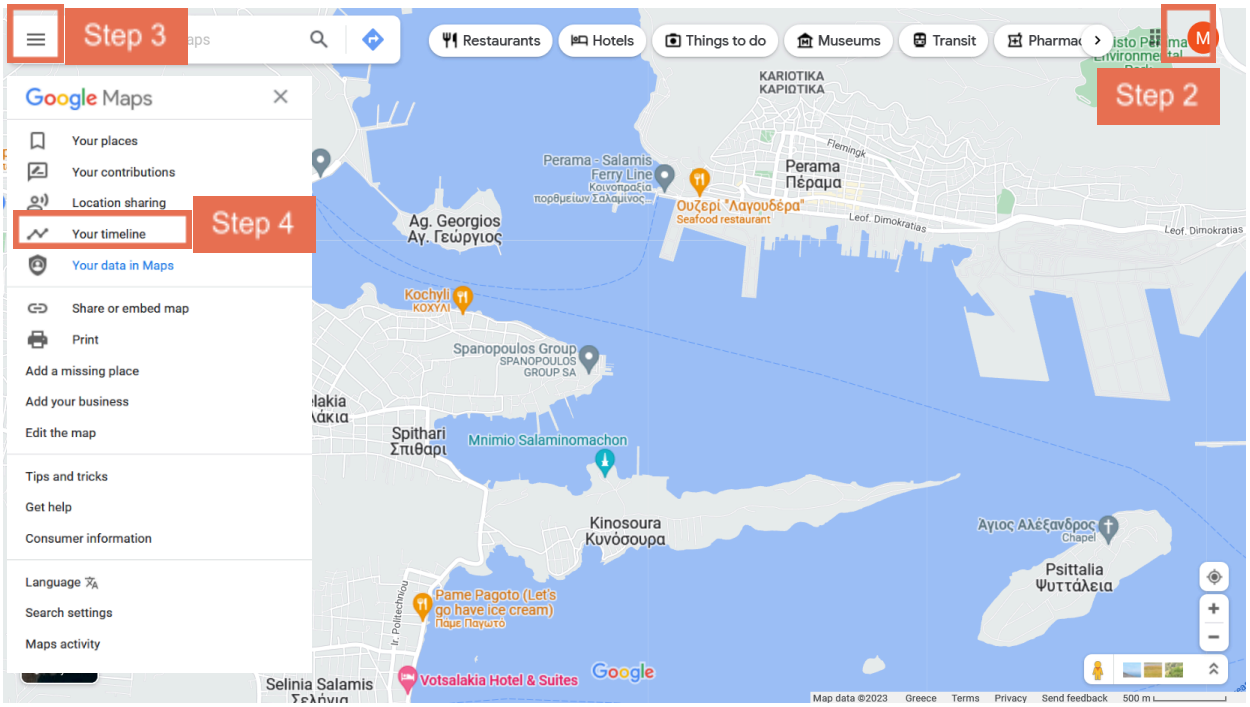
How to export Google Maps Timeline Data

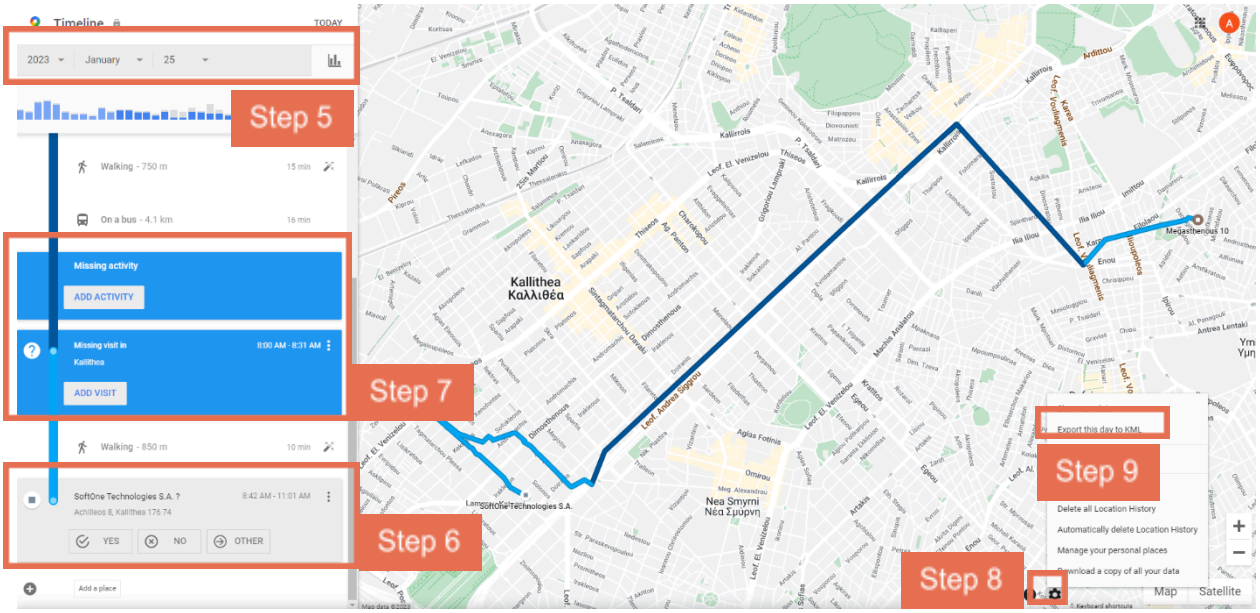
In order to export and share your Google timeline with TANGENT, please follow the following steps. *Note that you can only export your Timeline from your Computer.*

1. On your computer, visit Google Maps
2. If needed, at the top right, sign in to your Google Account
3. At the top left, click Menu

4. Click: Your Timeline
5. Select a typical day **from 2023**.
6. If prompted to, confirm the destination and the mode of your trips.
7. Consider adding information on your activities and visits.
8. Click the settings button at the bottom right.
9. Select “Export this day to KML”
10. Repeat the process for all traces of a typical month (**30 days**).
11. Create a zip file for all 30 KML files.
12. Send the files to: surveys.tangent@gmail.com

Visual Guide:





For more information on enabling, reviewing & editing Google Maps Timeline, visit the dedicated Google website:

Android iPhone & iPad Computer

TANGENT

Informed Consent

Travel Behavior Survey

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 955273

Your rights as participant

Participation in this study is voluntary and no costs are derived for you. You have the right not to participate at all or to leave the study at any time. Deciding not to participate or choosing to leave the study will not result in any penalty or loss of benefits to which you are entitled.

In case you want to withdraw, you can contact Eleni Vlahogianni (elenivl@central.ntua.gr) and all your data removed from the study and deleted.

Data protection and confidentiality

It is very important to us to protect your privacy when processing your data.

The information provided will be used only for the purpose of research. Your information will be assigned a code number that is unique to this study. The list connecting your name to this number will be kept in a locked file in the servers of the National Technical University of Athens and only the responsible of the investigation (NTUA researcher Eleni Mantouka) will be able to access the matching. When the study is completed and the data have been analysed, the list linking participant's names to study numbers will be destroyed.

Data will be collected in four different countries (UK, Greece, Portugal and France) and potentially in other EU countries as well, and transferred to Greece where it will be stored in NTUA's servers.

The traces collected will be stored in NTUA's servers and will be kept until 5 years after the end of the project (end date: 31 August 2024) for auditing or reporting purposes by request of the funding agencies. The data will be managed according to NTUA's Privacy Policy. No other disclosure will be done during this period. It will be deleted afterwards. The data will be used only for the purposes of the project and will be deleted afterwards.

At any moment you can contact Eleni Vlahogianni (elenivl@central.ntua.gr) to modify or delete your personal information.

Legal information

Data will be handled complying with the applicable regulation at EU level, as well as the national laws applicable:

At the EU level:

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (OJ L 119, 4.5.2016, p. 1).

For the United Kingdom:

UK - Data Protection Act 2018

Further information

If you have further questions or concerns about your rights as a participant in this study, please contact Eleni Vlahogianni (elenivl@central.ntua.gr), or the project coordinator Leire Serrano (leire.serrano@deusto.es).

You may submit complaints or grievances about the conduct of the project to the NTUA's Ethics Committee of Research (ereyna@central.ntua.gr) and complaints or grievances about the handling of your personal data with the NTUA's Data Protection Officer (elke_dpo@mail.ntua.gr) and the Hellenic Data Protection Authority (complaints@dpa.gr).

(Please continue to the next page)

Informed consent

By participating in this study, you explicitly agree with the following:

- Your participation in the action is voluntary.
- You are at least 18 years of age.
- You have read the TANGENT information sheet that provides enough details about the project (purpose, expected duration and procedures of the activity).
- You have been informed about your right to refuse to participate or to leave the activity at any moment without any justification.
- You have been notified of the contact persons, in the case you have questions or doubts during the activity.
- You have been informed that a copy of the consent form will be sent to your email account.
- You have had enough time to decide on your participation in the action.
- You have been informed about the storage procedures of the study data.
- You have been informed about the confidentiality of your personal data.
- You allow experts involved in the study under confidentiality agreements to utilize the information for the purpose of the action and only for this
- You agree to participate in the action.