National Technical University of Athens (NTUA)
School of Civil Engineering

Dynamic transport mode estimation subject to joint decisions and spatio-temporal variations

PhD Thesis | English Edition

Konstantinos Gkiotsalitis

July 2017

Supervisor:
Antony Stahopoulos, Emeritus Professor, NTUA

Advisory Committee:
Constantinos Antoniou, Professor, TU Munich
Athanasios Ballis, Associate Professor, NTUA

"The only impossible journey is the one you never begin"

Preface

This thesis studies the dynamic route and transport mode optimization for participating in a jointly decided activity subject to spatio-temporal variations. Joint activity participants have to travel from their current locations to a common location which is the location of the joint activity. Apart from the recurrent joint activities (such as work, school etc.), there are several joint leisure activities where activity participants have to decide about the location and the starting time of the joint leisure activity along with the transport mode(s) that each one of them has to use for commuting from his/her current location to the joint leisure activity location.

The objective of this thesis is the development of a comprehensive system for the optimization of joint leisure trips that are not related to work by optimizing (a) the location and the starting time of the joint leisure activity (b) the public transport operations (c) the transport mode(s) selection for each activity participant in order to arrive there as fast as possible while satisfying his/her personal trip preferences.

Activities not related to work can be responsible for more than 60% of trips at an urban environment. Non-working travel patterns differ from the more stable, recurrent travel patterns of work-related activities such as trips from/to work, school etc. and the three main differences of those activities are:

- The location of a leisure activity can differ on a daily basis (it is not static like the location of the working or studying place)
- The starting time of a leisure activity has greater elasticity and can differ on a daily basis (it is not stable such as the starting time of work, school etc.)
- The alternative journey options to and from a leisure activity location are not well-known to the users (users are more aware of their journey alternatives when it comes to transfers from/to work-related activities since those activities are re-current)

Given that a significant number of transfers is related to leisure activities, the optimization of the (1) location selection, (2) starting activity time, (3) transport mode selection and (4) route selection are of paramount importance for both the commuters' total travel cost and the transport network performance. In addition, the prediction of non-recurrent activities in time and space can be an important step forward for the tactical and dynamic planning of transport networks since the volume and the non-recurrent nature of such activities lead to significant travel demand variations compared to the more stable, work-related activities.

Due to the above, this thesis focuses on: (i) Understanding the State-of-the-Art (SoA) work on utilizing user-generated data for increasing the efficiency level of joint leisure activities and proposing actions towards this direction; (ii) Capturing users' willingness to travel certain distances for participating in different types of activities; (iii) Optimizing the selection of locations and starting times of joint leisure activities (iv) Re-scheduling the starting times of public transportation trips in order to adjust to the joint leisure activity demand without deteriorating the Quality of Service (QoS) for other passengers; (v) Optimizing the journey/path selection of users' who are willing to travel from one point of the network to another for participating at one activity and, possibly, utilize multiple modes while also satisfying their preferences.

Keywords: Social Media; Social Networks; Utility-Maximization; Joint Activities; Spatio-Temporal Pattern Recognition; Joint Travel Optimization; Stochastic Search; Regularity-based Public Transportation; Nonlinear Programming; Real-Time Multimodal Fastest Paths; Intermodal Networks; Resource Constrained Shortest Path Problem

- Konstantinos Gkiotsalitis

Detailed Contents

Figures

Tables

1. Introduction

1.1. Objective

The objective of this thesis is the development of a comprehensive system for the optimization of joint trips that are related to leisure (non-work) activities. Trip attractor locations which are not related to work activities can be responsible for more than 60% of trips at an urban environment. Non-work travel patterns do not exhibit the re-current nature of travel patterns related to work activities such as trips from/to work, school etc. The three main differences of non-work related trips are:

- The location of non-work activities can differ on a daily basis (it is not static like the location of the working or studying place)
- The starting time of the activity can differ (it is not stable such as the starting time of work, school etc.)
- The alternative transfer options to and from the activity are not well-known to the users (users are more aware of their transfer alternatives when it comes to transfers from/to work-related activities since those activities are re-current)

Given that a significant number of transfers is related to leisure activities, the optimization of the (1) location selection, (2) starting activity time, (3) transport mode selection and (4) route selection are of paramount importance for both the commuters' travel cost and the transport network performance. In addition, the prediction of non-recurrent activities in time and space can be a key enabler for the tactical and dynamic planning of transport networks. This is especially important since the volume and the non-recurrent nature of such activities lead to significant travel demand variations compared to the more stable, work-related activities.

The complexity of the optimization of leisure activities is an effect of the strong spatio-temporal variation of such activities; something, that requires continuous analysis of user-generated data related to commuters' trips instead of the use of more static approaches such as revealed preference surveys. For this reason, the development of comprehensive analytic tools for data mining and analysis of continuously updated user-generated data from social media and smartphones is required. The scope of data collection from private users is strictly the identification of commuters' transfers in space (origins/destinations) and time with the utilization of GPS-enabled smartphones and other smart devices.

One more issue is the process of large amounts of heterogeneous user-generated data and the quasi real-time optimization requirement regarding the selection of the leisure activity location, starting time of the activity, transportation modes and associated routes. This challenge requires the development of hybrid models and heuristic search methods for performing the optimization tasks with high accuracy and low computational costs. Finally, the user preferences of each commuter perplexes further the aforementioned problem requiring the inclusion of the perceived utility of each commuter to the analysis of the problem.

In the context of this thesis, the above issues are analyzed by focusing on:

- The research of optimizing mobility transfers related to joint leisure activities, the parameterization of this problem and the analysis of its characteristics
- The research of the mobility transfer options of commuters with the use of continuously updated user-generated data
- The development of algorithms for the selection of an optimal leisure activity starting time and location together with the selection of transport modes and routes for the associated mobility transfers
- The implementation of the developed solution methods at urban networks and the prediction of non re-current activities for the dynamic planning of public transportation

Those inter-dependent research activities are published in several peer-reviewed transportation journals with the aim of developing one comprehensive system for analyzing mobility transfers with significant spatio-temporal variations and optimizing (i) the travel costs related to commuters; (ii) the performance of the transport network; and (iii) the ridership levels of public transportation with demand-responsive dispatching.

1.2. Featured Publications

Parts of this thesis' chapters are related to the featured publications:

- Publication 1: A comprehensive literature review is provided and the unresolved issues that are analyzed in this thesis are underlined.
- Publication 2: Probabilistic models are developed for capturing commuters' transfer decisions based on user-generated data from Social Media. The input of those models is the location of the user and the associated timestamp of that location (GPS trace). The same input can be provided from Automated Fare Collection (AFC) data from public transport systems.
- Publication 3: Hybrid algorithms of high accuracy are developed for optimizing the selection of location and starting time of joint leisure activities with computational complexity that allows quasi-real-time applications.
- Publication 4: The problem of dynamic re-planning of public transport schedules (especially bus operations) for increasing the ridership related to non-recurrent activities with significant

spatio-temporal variations is modeled and optimized with the introduction of evolutionary optimization techniques.

- Publication 5: Resource-constraint path selection algorithms are developed for optimizing the journeys of users' who are willing to travel from one point of the network to another with the use of one or multiple modes while also satisfying their preferences.

Publications

1. K. Gkiotsalitis, A. Stathopoulos (2015): "Optimizing Leisure Travel: Is BigData Ready to Improve the Joint Leisure Activities Efficiency?", vol.38 of the series Engineering and Applied Sciences Optimization, Springer International Publishing, 53-71.
2. K. Gkiotsalitis, A. Stathopoulos (2015): "A utility-maximization model for retrieving users' willingness to travel for participating in activities from big-data", Transportation Research Part C: Emerging Technologies, 58, 265-277.
3. K. Gkiotsalitis, A. Stathopoulos (2016): "Joint leisure travel optimization with user-generated data via perceived utility maximization", Transportation Research Part C: Emerging Technologies, 68, 532-548.
4. K. Gkiotsalitis, A. Stathopoulos (2016): "Demand-Responsive Public Transportation Re-scheduling for adjusting to the Joint Leisure Activity Demand", International Journal of Transportation Science and Technology (IJTST), Elsevier, vol. 5, issue 2, 68-82.
5. K. Gkiotsalitis, A. Stathopoulos (2015): "A mobile application for real-time multimodal routing under a set of users' preferences", Journal of Intelligent Transportation Systems 19 (2), 149-166.

2. Background Theory and Methodology

2.1. Literature Review

Introduction to non-recurrent activities

Today's metropolises with complex transport networks and numerous places for leisure activities pose great challenges to individuals who are willing to organize and participate in joint activities with non-recurrent nature (i.e., not work/school related activities). In this study, we consider as joint leisure activities all activities conducted out-of-work involving the participation of two or more travelers.

[1] posed that 29.2% of all daily trips are related to leisure activities, while 28% are conducted for shopping and personal business and 10.7% for other activities including escort. Similar results were observed on the New York Regional Travel survey [2] . Given the surveys' insights, it is evident that almost 70% of all conducted trips (*typical weekday trips* = 2.51 × *number of inhabitants* in the city of London) have a non-recurrent nature and can be potentially shifted to the category of leisure joint trips. The aforementioned action is expected to promote the interpersonal relations among individuals via increasing the number of physical meetings and improving the planning efficiency of out-of-work activities.

[1]: (2014), *Transport for London, Travel in London, London Travel Demand Survey*

[2]: (2014), *New York Regional Travel Survey*

Fixed trips with recurrent characteristics (i.e., trips to work/school) can be recorded more easily enabling the central transport authority to act on smoothing congestion and the individual agent to act on reducing his/her travel costs via changing the transport/working schedules or shifting the departure times respectively. In contrary, leisure trips have a non-recurrent nature which complicates the implementation of control actions for the relief of congestion.

In the case of recurrent trips, travelers observe the repeated congestion patterns since they keep experiencing them on a daily basis while traveling over similar origin/destination points. Therefore, they have enough information for adapting their schedules in order to reduce their waiting times. For instance, travelers are well-informed regarding the traffic conditions for trips to/from work due to their prior experience on traversing the same path on a daily basis, while they are less aware of the feasible set of trip-selection options when planning their out-of-work or other non-recurrent trips. Consequently, the lack of information yields three main inefficiencies:

- Fluctuation of Travel Demand: Out-of-work trips cannot be easily predicted from the transport network operator due to the significant day-to-day variation

- Interpersonal Activities Loss: Not aware of the daily schedules of other individuals, one examined agent is either not able to schedule a joint leisure activity or schedules an inefficient one with high opportunity cost and limited participants
- Trip Selection Inefficiency: Individuals enumerate a number of possible trips and select a most-preferred option via simple permutation or perceived utility-maximization without being perfectly informed during the decision-making process

At this point, it should be stated that the individual-level planning of trips in metropolitan areas cannot be perceived as fully inefficient since it is based on a perceived utility-maximization approach; however, the lack of perfect information on the decision-making phase affects the efficiency of trip selection. Failure to construct a utility function which corresponds to the real-world conditions leads to the maximization of an ill-defined problem.

The utility function is perceived correctly if the individual is well-informed during the trip-selection via holding information over three separate dimensions (refer to Fig.2.1):

- The current traffic on the road network and delays on public transport services
- The exact location of all places of interest for leisure activities in the examined metropolis (i.e., location of bars, restaurants, cinemas)
- The daily schedules and the preferences of all friends and acquaintances of an individual with whom a joint leisure activity can be organized (the degrees of freedom might differ depending on the social network of the examined individual)

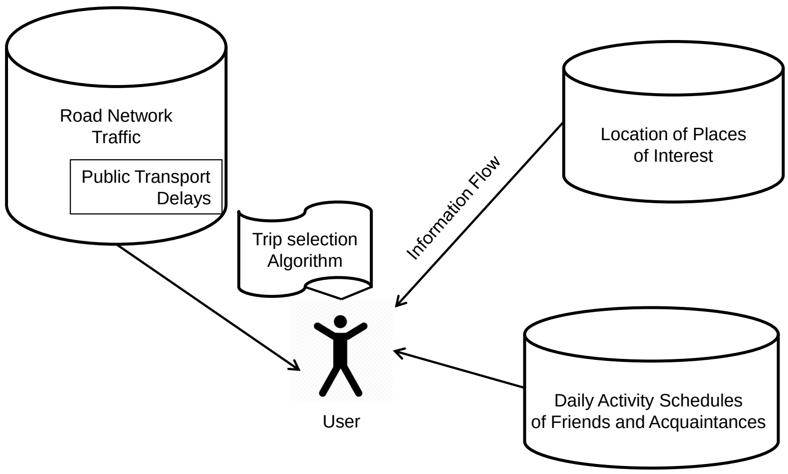


FIGURE 2.1.: The three dimensions of Information Flow for Trip selection at the individual level

[3]: Gonzàlez et al. (2008), 'Understanding individual human mobility patterns'

Several attempts have been made to define special laws for modeling the movement of people (refer to [3]). However, the lack of information at the trip-selection phase hinders the maximization of utility. At this stage, the utilization of new information streams can be seen as a valuable resource for improving the awareness over the three dimensions of the decision-making process. In an example of improving a service via raising the level of information dissemination, early research based on bus riders demonstrated various positive effects such as increased ridership and traveler satisfaction attributed to enhanced information availability. That showed that easy access to relevant travel information is a decisive factor for the success and adoption of public transport systems (refer to [4]).

[4]: Ferris et al. (2010), 'OneBusAway: Results from Providing Real-Time Arrival Information for Public Transit'

Given the above, the state-of-the-art in the area of non-recurrent trips is examined with the aim of improving the mobility efficiency related to such activities with the use of insights from user-generated data. Attention is given to studies on utilizing near real-time user-generated data (i.e., data from smartphones, smartcards, PDAs) for tackling transportation problems. The aim is to formulate the problem of optimizing leisure travel considering the provision of user-generated data, present the direction of the state-of-the-art, understand why user-generated data has not been used for increasing the volume and the efficiency of joint leisure activities and propose actions to move towards this direction.

Utilizing Cellular Data in Transportation Problems

Cellular data is the form of user-generated data which have been studied the most for predicting individuals' mobility patterns even if the positioning via cell towers is not as accurate as the GPS positioning. Regardless the posed challenges, cellular data have been utilized to improve the understanding of human mobility and develop individual-level models for capturing the mobility and activity habits of individuals. Since cellular networks need to know the approximate location of all active phones to provide them voice and data services, location information from Call Detail Records (CDRs) can be used for characterizing the mobility patterns of hundreds of thousands of people.

The most common individual-level models for predicting the mobility of individuals -which are not based solely on spatio-temporal travel pattern recognition- are the activity-based models (refer to [5] , [6], [7], [8]). Those models are the basis for forecasting individuals' daily trip schedules from cellular data and perceive each trip as a mean to participate at pre-scheduled activities.

[5]: Axhausen et al. (1992), 'Activity-based approaches to travel analysis'

In the literature, [9] utilized Cellular data logs for correlating the mobility patterns of an individual with the mobility patterns of his friends and acquaintances. The underpinning theory of the correlation process includes the assumption that users' travel patterns do not depend on time and space, but also on the travel patterns of other individuals inside their social network. The findings of the research showed that the mobility patterns of one examined agent can be predicted more accurately when the mobility patterns of his/her social network are considered as explanatory variables. [10] worked also on the same direction using data from the Nokia Mobile Data Challenge dataset. The work of [9] can provide some evidence on the theoretical concepts developed by [11], [12] and [13] on predicting agents' mobility based on their social networks. Those theoretical concepts place the traveler at the center of decision-making (ego-centric approaches) and

[9]: Musolesi et al. (2007), 'Designing Mobility Models Based on Social Network Theory'

offer a new framework for microsimulation where harvesting large-scale user-generated data is expected to facilitate the implementation phase.

[14]: Calabrese et al. (2011), 'Estimating Origin-Destination Flows Using Mobile Phone Location Data'

In addition, [14] , [15], [16] and [3] utilized cellular data for predicting the mobility patterns of individuals in urban scenarios over time and space. Those studies, including studies of [17], [18], [19] and [20], attempted to exploit the emergence of the mobile positioning technology and the market penetration of mobile phones by developing methods for estimating the OD matrices in study areas. In the same way, [21] used the cellular communication system and cell phone towers to transfer information and estimate OD matrices.

[21]: Sohn et al. (2008), 'Dynamic origin–destination flow estimation using cellular communication system'

In a practical example, in the work of [14], an algorithm for estimating a population's travel demand in terms of ODs from aggregating the trips of individual mobile phone users in the Boston Metropolitan area was developed. During the validation, it was shown also that the estimated OD flows were close to the US Census estimates. [22] proceeded further on comparing the mobile-phone-based mobility traces against data from odometer readings from the annual safety inspections in the Boston Metropolitan area to check the validity of mobile phone data in characterizing individual mobility and to identify the differences between individual mobility and vehicular mobility.

The limits of predictability in human dynamics by analyzing mobility patterns of mobile phone users were also analyzed and evaluated by [23] . More recently, [24] and [25] proposed a methodology for using mobile phone data to analyze the mechanism of trip generation, trip attraction and the OD information with a pilot study in Beijing via using the K-means clustering algorithm to divide the traffic zones. In addition, [26] worked on a method for separating trips (capturing the starting and ending points of a trip) on the basis of GPS data collected by smartphones after considering that even when the subject stays into a place, the collected GPS coordinates are not always exactly the same according to the surrounding environment (precision is 81% and recall rate is 62%). Apart from detecting departure and arrival times, methods for classifying automatically modes of transportation on the basis of smartphone GPS data were also proposed by [27].

[23]: Song et al. (2010), 'Limits of Predictability in Human Mobility'

Finally, in another set of studies from [28] and [29], Bluetooth devices were distributed to people to collect mobility data and study the characteristics of co-location patterns among people.

[28]: Hui et al. (2005), 'Pocket Switched Networks and Human Mobility in Conference Environments'

To summarize, works on utilizing cellular data in transportation have been focused on different areas:

- Estimating the OD matrix in a study area
- Exploring the mobility patterns of one individual based on the mobility patterns of his/her social network
- Extracting the current mode of transportation
- Extracting the starting and ending time of a trip

However, there are no works in the area of activity-participation analysis which can facilitate the development of new applications for suggesting common activities to users with social ties based on their willingness to participate simultaneously in similar activities being held in nearby locations.

User-generated data from Social Media and its applications in transportation

The research on data from social networks on understanding users' mobility is in its early stages. The first studies focused on the power of micro-blogging on offering near real-time insights on crisis events when all other means of communication have failed. Routinely, the importance and the volume of the crisis event is captured through the magnitude of micro-blogging messages and their content information. A study from [30] explored crisis informatics using Twitter data after the Oklahoma Moore tornado demonstrating the potential of social media data on extracting relevant information during natural disasters.

[30]: Ukkusuri et al. (2014), 'Exploring Crisis Informatics Using Social Media Data: A Study on 2013 Oklahoma Tornado'

In a similar fashion, social media data from social networks like Facebook, Twitter, Foursquare and the image sharing service, Flickr, have already been used in research works describing crisis or natural disasters such as Virginia Tech shooting ([31]), Southern California wildfires ([32]), major Earthquakes in China ([33], [34]), Red River floods and Oklahoma grassfires ([35]).

[32]: Hughes et al. (2008), 'Sight-Seeing in Disaster: An Examination of On-Line Social Convergence'

In another set of works, [36] utilized the Internet as resource to capture the crowd levels during planned special events. In general, local events are not tracked from transport authorities since manual, labor-intensive tracking is needed. [36] utilized the Internet as a resource for contextual information about special events and developed a model that predicts public transport arrivals in event areas. The results were demonstrated with a case study from the city-state of Singapore using public transport tap-in/tap-out data coupled with local event information obtained from the Internet after performing primitive data fusion.

[36]: Pereira et al. (2013), 'Using data from the web to predict public transport arrivals under special events scenarios'

In another work, [37] focused on developing and testing analytic techniques for fusing user-generated data from Social Media and smart-cards in order to capture the mobility patterns in urban areas. Automated methods for retrieving users' mobility patterns from historic, user-generated data logs, comparing user' profiles based on the similarity of their observed mobility patterns and categorizing users in clusters were developed. During the testing phase, user-generated data from London Smart Card and Social Media users were utilized to cluster users based on their mobility-activity pattern similarities. Results showed that it is possible to integrate data logs from multiple sources to capture the main mobility-activity patterns observed in an area. However the topic of joint participation in non-recurrent activities has not been addressed until now.

[37]: Gkiotsalitis et al. (2014), 'Mobility Demand Prediction in Urban Scenarios through Multi-source, User-generated Data'

Social media have also been used for capturing the activity types performed by users at different locations via advanced spatio-temporal analysis and educated rules (refer to [38]). In the same work, techniques for estimating individuals' daily schedules and the sequence of

[38]: Gkiotsalitis et al. (2014), 'Educated Rules for the Prediction of Human Mobility Patterns based on Sparse Social Media and Mobile Phone Data'

activities were developed. [39] focused also on the same topic introducing a probabilistic model for modeling individuals' daily schedules based on input data from several sources (i.e., Social Media, Cellular Data). Along the same lines, new techniques to detect urban mobility patterns and anomalies by analyzing trajectories mined from publicly available geo-positioned tweets were presented by [40].

A set of recent methods for capturing the geographic user activity patterns based on Foursquare have also been developed in the works of [41] , [42], [43] and [44].

[41]: Noulas et al. (2011), 'An Empirical Study of Geographic User Activity Patterns in Foursquare.'

Summarizing, social media data which is individualistic in nature has been utilized for:

- Capturing the volume and the effects of crisis events
- Estimating individuals' mobility patterns and correlating them based on observed patterns from other datasets
- Retrieving activity types of users
- Capturing the arrival times and the expected demand at local events

User-generated data from Smart Cards

With the deployment of automatic fare collection systems, large-scale data becomes available for real-world transport usage [45] . As more and more sensors have been integrated into public transport infrastructures, large-scale transport data is produced at high rates [46]. Nonetheless, studies of estimating individual travel patterns with smartcard data are sparse in public transport research compared to studies on cellular data and social media.

[45]: Pelletier et al. (2011), 'Smart card data use in public transit: A literature review'

In the past, research has mainly focused on aggregate demand forecast [47] . Based on a gravity model, [48] showed that some of the variation in mobility flows is influenced by distance and population of local residents via analyzing smartcard data, while [49] analyzed time series of AFC data to identify events of overcrowding at public transport stations. [50] and [51] also measured the transit use variability with smart-card data.

[47]: Chatterjee et al. (2011), *Travel Demand Forecasting for Urban Transportation Planning*

The potential of smart card data for travel behavior analysis in Britain was studied by [52] where the pensioner concessionary pass in Southport, Merseyside, and the commercially operated scheme in Bradford were examined. There was stated that the nature of smart card data puts an emphasis on concept definition and rule-based processing; but limitations were also recognized such as the trip lengths which are not recorded on systems. The latter implicates also the efforts on performing individual-level analysis and predicting individuals' daily travel schedules.

[53] addressed the above-mentioned issue by measuring the variability of mobility patterns after analyzing a full-week smart-card dataset from Singapore. In spite the limited study period, it was possible to demonstrate that the number of trips and mobility patterns vary from day to day but the overall spatial structure of urban movements

[53]: Zhong et al. (2015), 'Measuring variability of mobility patterns from multiday smart-card data'

remained the same throughout the week. However, this finding cannot be generalized due to the limited period of the analysis (only one week).

[54] utilized travel card data from a large population of bus riders from Lisbon, Portugal. The main intention of the work was to predict the future bus stops accessed by individual drivers and it was demonstrated that accurate predictions can be delivered by combining knowledge from personal ride histories and the mobility patterns of other riders. In another work, [55] utilized smart card data from a bus line in Singapore for developing a modeling platform for testing bus transportation.

[54]: Foell et al. (2014), 'Catch me if you can: Predicting Mobility Patterns of Public Transport Users'

[55]: Ivanchev et al. (2014), 'Stochastic Bus Traffic Modelling and Validation Using Smart Card Fare collection data'

Several research works such as [56] and [57] had a special focus on collecting spatial travel information provided by AFC systems for estimating Origin-Destination Matrices from smart card data; whereas, [58] focused also on the high-level analysis of travel behavior in Singapore with the use of one-day EZ-link data.

[56]: Li et al. (2011), 'Estimating a transit passenger trip origin-destination matrix using automatic fare collection system'

Finally, as discussed before, in the work of [37] a more individual-based approach was considered for identifying users' mobility patterns from historic smart card data logs. For that case study, data from 200 Oyster card users in London were utilized.

[37]: Gkiotsalitis et al. (2014), 'Mobility Demand Prediction in Urban Scenarios through Multi-source, User-generated Data'

It is evident that smartcard data offers less qualitative information compared to social media or cellular data generating problems on predicting the daily schedules and the social networks of individuals since it represents only public transport mode passengers.

Use of GPS traces from personal navigators and smartphones in transportation

Recent research studies have also focused on the utilization of user-generated global positioning system traces generated from personal navigators and smartphones in the transportation domain. For instance, applications request from the user of a GPS-enabled device to record his/her geo-location of any record travel/sports experience, and then upload, visualize and browse their GPS data on a Web map (specific implementations at [59] , [60], [61]). Further, users are enabled to exchange life experiences by sharing GPS logs in the Web community. [62] attempted to estimate the activity patterns of smartphone users based on GPS tracking data. They developed a method for classifying "indoor", "outdoor static", "outdoor walking", and "in-vehicle" status. Similarly, [63] developed a special device, called a behavioral context addressable logger (BCALs), for collecting various kinds of data such as GPS coordinates, acceleration, atmospheric pressure, angular velocity, UV index, direction, and loudness. BCALs can distinguish situations in which smartphone users are classified as "walking", "up/down-staircase", "bicycling", and "in-store".

[59]: Krumm et al. (2007), 'Predestination: Where do you want to go today?'

[63]: Hato (2010), 'Development of behavioral context addressable loggers in the shell for travel-activity analysis'

There are also studies on trip-separation methods (mainly by utilizing a series of user-generated GPS-traces from GPS-enabled devices for capturing the starting and ending time of a trip) from [64] , [65], [66],

[64]: Li et al. (2008), 'Mining user similarity based on location history'

[67]. For instance, [64] utilized users' location histories from their GPS-enabled devices for implying their interests and deriving the level of similarity among users based on their geo-location histories. Among those studies, only [68] reported an evaluation of a trip-separation method. The basic idea forming the basis of their method is to find the so-called "stay points". They regard consecutive GPS coordinates as stay points if they satisfy the following two conditions: (i) they fall within a circle with diameter of 196 meters; and (ii) the time difference between the first and last stay points is more than 14 minutes. The key idea behind this stay-point detection is the elimination of outliers that can cause mis-detection. This trip-separation method achieved precision of 92.4% and recall rate of 90.5%.

[68]: Witayangkurn et al. (2013), 'Trip Reconstruction and Transportation Mode Extraction on Low Data Rate GPS Data from Mobile Phone'

Given the above, one can conclude that geo-location data from smartphones or personal navigators have been mainly utilized for:

- Estimating OD matrices
- Capturing the type of utilized transportation
- Activity-pattern estimation
- Separation of trips

Discussion on Background Theory

Studying the background theory was an attempt to investigate how different forms of user-generated data (cellular, social media, smart card and personal navigator data) have been utilized so far. The objective was to examine if the data sources and the developed techniques have some potential on increasing the efficiency of joint leisure activities in today's metropolis.

In a first attempt to summarize the results, Table 2.1 provides aggregated information on the utilization of user-generated data from different sources according to the state-of-the-art studies.

Table 2.1.: Aggregated information on the utilization of user-generated data from different sources according to the state-of-the-art studies

	Cellular	Social Media	Smart Card	GPS Positioning
Estimating OD matrices	Yes	No	Yes	No
Extracting the utilized mode of transportation	Yes	No	Yes	No
Capturing the Trip Separation	No	No	Yes	Yes
Real-time Traffic estimation	Yes	No	No	Yes
Estimating the daily schedule of agent's social network	No	Yes	No	No
Crisis Events Analysis	No	Yes	No	No
Capturing Individuals' Mobility Patterns	Yes	Yes	No	No
Retrieving the performed activities by users	Yes	Yes	No	No
Forecasting the expected demand at local events	No	Yes	No	No
Separating Trips	Yes	No	No	Yes
Activity-pattern estimation	Yes	Yes	Yes	Yes

During the exploration of non-recurrent activities, three information dimensions were considered for assuming that an individual is perfectly informed for making an optimal decision on selecting a leisure joint activity. In Table 2.2, it is shown which kind of information is expected to be retrieved from different sources of user-generated data. From Tables 2.2, 2.1 one can observe that although the full information for forming a decision-making objective function is obtainable, research works have not been focused on that direction.

Table 2.2.: Potential of user-generated data on providing information for joint leisure activity planning

	Cellular	Social Media	Smart Card	GPS Positioning
Real-time traffic	No	Yes	Yes	Yes
Real-time public transportation arrival times at stops	No	Yes	Yes	Yes
Location of Places of Interest	Yes (with low accuracy)	Yes	No	Yes
Daily Schedules of agent's Social Network	No	Yes	No	Yes

Due to the above, the importance of developing new models for tapping the potential of user-generated data for improving the efficiency of joint leisure activity planning is highlighted. New models and techniques are recommended to focus on the following:

– Data processing tools

- Algorithmic tools for data aggregation and fusion
- Processing tools that can calculate the optimal point(s) of the utility function and return an optimal joint leisure activity to the traveler

Proceeding towards this direction, the aim is to influence the ~70% of out-of-work trips in metropolitan areas for improving the trip planning and operations; thus, yielding significant gains for both the individual traveler and the central transport authorities.

2.2. Data Processing and Methodology

Discussion on Methodology

To improve the joint leisure activity trips in metropolitan areas, a set of different problems are considered that focus on the (1) location selection, (2) starting activity time, (3) transport mode selection and (4) route selection. The scope of this multi-stage approach is the reduction of the commuters' travel costs and the improvement of the transport network performance.

This multi-stage approach is detailed in the following chapters of this thesis when different methodological strategies are introduced. However, a brief overview of the introduced methodologies is provided in this section for offering a general guide for this thesis. This section presents the introduced methodologies at a higher level and justifies the content and the structure of the following chapters of this thesis that attempt to provide a holistic system for the optimization of joint leisure trips.

The basic pillars of this thesis are the methodologies related to (a) capturing users' willingness to travel certain distances for participating at different types of activities based on user-generated data; (b) optimizing the location and time of joint leisure activities; (c) adapting the public transportation timetables to the joint leisure activity demand and, finally, (d) planning multi-modal journeys for the joint leisure activity participants based on personalized preferences. Those methodologies are summarized in Fig.2.2 and each of the (a-d) topics are analyzed in the thesis chapters 3-6.

At first, the utilization of user-generated data for mobility purposes is analyzed at the literature review section considering user-generated datasets from Cellular Data, Social Media, Smart-Cards and GPS-enabled devices (i.e., personal navigators, smartphones).

In Chapter 3, methodologies are introduced for utilizing user-generated data in order to capture users' willingness to travel certain distances for participating at different types of activities. One of the introduced methods is a daily pattern recognition model that analyzes historical user-generated data and derives the daily mobility-activity patterns of individual users. For achieving this, a secondary model is introduced for associating re-visited locations with users' activities. Finally, individualistic utility-maximization models are introduced that capture

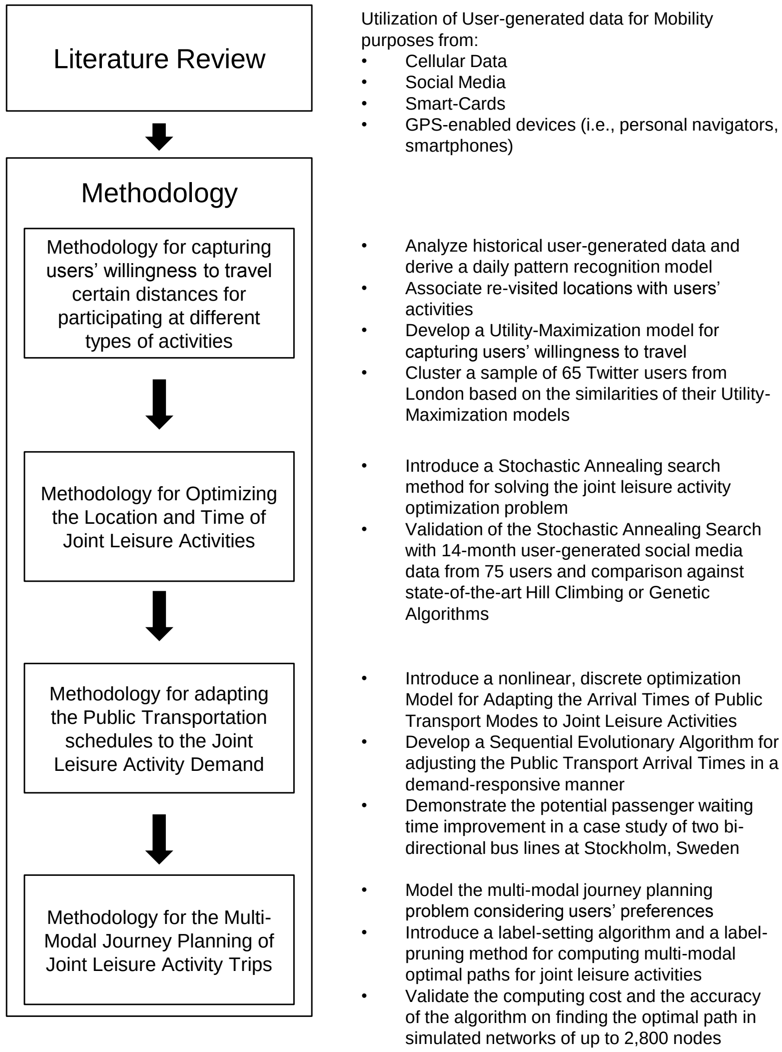


FIGURE 2.2.: Methodology of the Thesis

users' willingness to travel certain distances for participating at joint leisure activities based on their historical user-generated data. Those models are validated with the use of a sample of 65 Twitter users from London who shared their geo-location information while tweeting over a period of fourteen months.

In Chapter 4 the models that return the willingness of each single user to travel certain distances for participating at different types of joint leisure activities given the time of the day are utilized for optimizing the location and time of joint leisure activities. Given the complexity of such optimization at a metropolitan area with hundreds of alternative places of interest, a problem-specific stochastic annealing heuristic search is introduced and it is implemented in an expanded set of 75 users from London while a set of comparison tests against state-of-the-art Hill Climbing or Genetic Algorithms is performed.

After determining the optimal location and time of joint leisure activities, the effect on public transport services is analyzed. Attention is given to demand-responsive public transport services that can change their timetables, and more precisely the departure times of their trips, for adjusting to the joint leisure activity demand without deteriorating the level of service or increasing the total operational costs. This problem is modeled in Chapter 5 leading to a non-convex, discrete optimization problem subject to a number of operational constraints that is computationally intractable. For this reason, a sequential evolutionary algorithm is introduced and the departure times of the public transport services are adjusted to the joint leisure activities' locations and times in a heuristic manner. The potential improvement of joint leisure activity passengers' waiting times is showcased in a case study of two bi-directional bus lines in central Stockholm, Sweden.

Finally, the planning of multi-modal trips for the joint leisure activity participants based on personalized preferences is modeled in Chapter 6 with the use different layers that represent different modes of transport that potentially merge at interchange stops. Given the complexity of the multi-modal shortest path problem under users' preferences, the problem formulation transforms users' preferences to optimization constraints with associated threshold values for monitoring the constraints' satisfaction and a label-setting method is introduced. The label-setting method performs a shortest path heuristic search by exploring the combination of different modes of transportation; however, this method fails to scale in vast networks. Therefore, a label-pruning method is introduced that works together with the label-setting method for searching more efficiently the solution space. To validate the computing cost and the accuracy of this methodology, optimal multi-modal paths of users were computed on vast simulated networks of up to 2,800 nodes and 30,000 links that emulate potential networks of metropolitan areas.

Data Requirements and Data Processing

After presenting the methodologies that are described in detail in Chapters 3-6, the data requirements for applying those methodologies are analyzed. The main objectives are to indicate:

- which are the minimal data requirements for optimizing joint leisure activities and implementing the methodologies described in Chapters 3-6
- which are the potential data sources (including all alternative options)
- which data sources are used in this thesis and why those data sources are selected
- which were the obstacles on selecting the appropriate data
- how this data has been processed and what is the structure of the utilized datasets

The data requirements for optimizing joint leisure activities and implementing the methodologies described in Chapters 3-6 are presented in Table 2.3.

Table 2.3.: Data Requirements for a Holistic Use-Case Application

Capturing users' willingness to travel certain distances for participating at different types of activities	Historical User-generated Data of all persons in a study area (i.e., a city) containing their geo-location traces and the activities performed at each location
Optimizing the Location and Time of Joint Leisure Activities	Historical User-generated Data of all persons in a study area including information about their social networks (i.e., the list of friends and acquaintances of each user with whom he/she is probable to participate at joint leisure activities)
Adapting the departure times of Public Transportation services to the Joint Leisure Activity Demand	Real Time GTFS data of Public Transport Services and User-Generated Data of all persons in a study area as described in the two above-mentioned cases
Multi-Modal Journey Planning of Joint Leisure Activity Trips subject to personalized preferences	Topology of a study area. Travel times of all links in the study area for all alternative transport mode choices (i.e., bus, private vehicles etc.). Link traversing cost and fuel consumption of each link given the utilized transport mode. Preferences of users together with the origin-destination points of their trips

In addition, the data sources (and the potential alternative options) that can provide the required data for implementing the methodologies described in Chapters 3-6 in a metropolitan area are presented in Table 2.4.

Table 2.4.: Required Data Sources for a Holistic Use-Case Application

Capturing users' willingness to travel certain distances for participating at different types of activities	**Social Media Data** of all individuals in the study area including their geo-location changes or **Smartphone or personal navigator** geo-location data from all users in a study area
Optimizing the Location and Time of Joint Leisure Activities	**Social Media Data** of all individuals in the study area including their geo-location changes
Adapting the departure times of Public Transportation services to the Joint Leisure Activity Demand	**GTFS** data of Public Transport Services and their operational constraints and **Social Media Data**, user-generated Data of all persons in a study area
Multi-Modal Journey Planning of Joint Leisure Activity Trips subject to personalized preferences	Topology of a study area. Travel times of all links in the study area for all alternative transport mode choices (i.e., bus, private vehicles etc.). Link traversing cost and fuel consumption of each link given the utilized transport mode. Preferences of users together with the origin-destination points of their trips

Finally, the utilized data in this thesis is presented in Table 2.5 which is followed by a discussion on why this data is utilized.

Detailing further the data requirements for each chapter, in chapter 3 the main scope is to capture users' willingness to travel certain distances for participating at different types of activities. As presented in Fig.2.4, Social Media data of all individuals in one study area including their geo-location changes or Smartphone/personal navigator

Table 2.5.: Utilized Data Sources in this thesis for Methodological Validation

Capturing users' willingness to travel certain distances for participating at different types of activities	user-generated tweets with geo-tagged locations from 65 twitter users in London over a 14-month period
Optimizing the Location and Time of Joint Leisure Activities	user-generated tweets with geo-tagged locations from 75 twitter users in London over a 14-month period
Adapting the departure times of Public Transportation services to the Joint Leisure Activity Demand	GTFS data from two bi-directional central bus lines in Stockholm, Sweden together with the list of operational constraints and **Social Media Data**, user-generated Data of 62 persons in the study area
Multi-Modal Journey Planning of Joint Leisure Activity Trips subject to personalized preferences	Topology of a study area. Travel times of all links in the study area for all alternative transport mode choices (i.e., bus, private vehicles etc.). Link traversing cost and fuel consumption of each link given the utilized transport mode. Preferences of users' undertaking trips together with their origin-destination points

geo-location data from all users is required. This user-generated data is required at the level of each individual for developing daily pattern recognition models for each user that derive the daily pattern mobility-activity patterns of individuals from their historical user-generated data. For this reason, the geo-location of each individual at each time where he/she travels between successive activities is required for understanding what was the origin of each trip and what the final destination. Such information-rich data that includes the geo-location of individuals can be provided from GPS-enabled devices that belong to those users. Those devices can be GPS-enabled personal navigators/smartphones that record any spatial movement of the smartphone holder or social media posts that contain information about the actual geo-location of the user when he/she posts. Especially for the social media posts, the user should utilize a GPS-enabled smartphone that shares his/her geo-location when he/she posts or should provide information regarding his/her location manually via location tags[1] .

1: a common practice on Facebook, Foursquare

This data should be retrieved from all users in a study area and, ideally, should include all their data logs over a considerable period of time (i.e., one year) for deriving their most common daily mobility/activity patterns and preferred traveled distances. Apart from those data sources, cellular data from all users in a study area can be utilized but offers lower accuracy regarding the locations of the users compared to GPS positioning; and thus cannot pinpoint the specific locations when a user performs certain activities. In addition, public transportation smartcard logs from users where the location of each tap-in and tap-out is recorded together with the ID of the user can be utilized. Nevertheless, if someone utilizes solely smartcard data in a case study and implements the methodologies described in this thesis he/she should be aware that he/she optimizes solely the joint leisure activity trips that are related with public transportation users and a large percentage of trips within the study area that are served from other means of transportation (including walking) are not considered.

At this point, it should be stated that it is very difficult in practice to acquire the high-granularity, user-generated data including all geo-location changes of all users within a study area over a significant time period (i.e., 1 year) based on the logs from GPS-enabled personal devices. At first, only a very small fraction of users in a study area is expected to use a personal navigator or a GPS-enabled smartphone that tracks his/her moves all the time. Therefore, the approach of utilizing posts from social media where individuals share also their location information is preferred.

In this thesis, we considered the most commonly used social media such as Facebook/Instagram and Twitter as potential input data sources for applying the methodologies of Chapter 3. Regarding Facebook, each user is obliged to use a location tag when he/she posts instead of providing his geo-location from a GPS-enabled device. More importantly, mining the posts of users on Facebook in a data crawling campaign requires the authentication token for each user's profile. This authentication token is private for each user and can be provided only upon request limiting the potential of retrieving all posts from all users of a study area. In contrary, posts on Twitter are publicly available (all posts from a user can be seen from all other users and there are no privacy restriction categories such as appearance of a post only in a closed group of friends etc.).

Due to the above, this thesis focused on data crawling campaigns on Twitter. The methodology for retrieving automatically users' willingness to travel certain distances for participating in different activity types over different days and times is tested for 65 twitter users residing in London after processing their publicly available, user-generated data including geo-tagged locations and time stamps of their interactions. The dataset contains a sum of 6,400 interactions collected through a crawling campaign from November 2012-January 2014. Ideally, all twitter users within the study area (London) should have been included in the data but in this thesis only 65 users were utilized. The reason behind this is that many twitter users post tweets without including their geo-tagged locations. Therefore, one cannot use those tweets to understand where the user is at the moment of tweeting and when he/she moves from one location to another to establish a set of mobility/activity patterns.

In practice, even if a large portion of Londoners tweet only a small fraction of them satisfied the two basic principles for the inclusion in the data sample. Those two principles are (1) the activeness (one individual should be an active twitter user with several tweets on an average day for at least a 4-month period for tracing his/her location at different times of the day and start developing his/her daily mobility/activity patterns) and (2) the geo-tagging of the posting location. Especially the last requirement of accepting only users with a vast majority of geo-tagged tweets limited a lot the number of Londoners which were included in the data sample because only a small portion of twitter users tweet from a GPS-enabled smartphone and permit the location of their tweet to be posted along with their tweet. The main reason is that each twitter user has to navigate his twitter account and select manually the option that permits the sharing of the geo-tagged

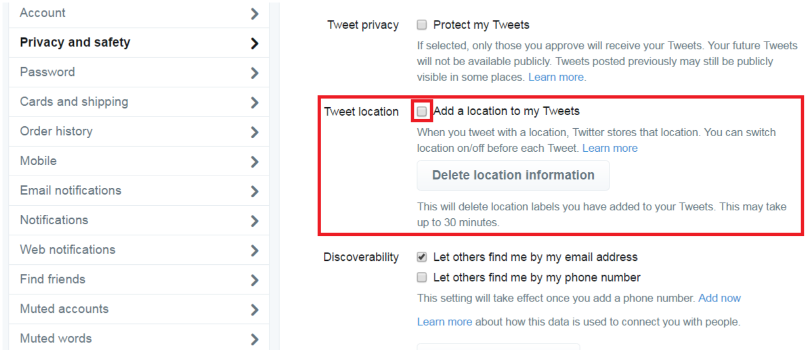


FIGURE 2.3.: Users can share publicly the geo-tagged location of their tweet when utilizing a GPS-enabled smartphone by selecting "Add a location to my tweets"

location of his/her tweets in public which is not common because most users are unaware of that option or are concerned about their privacy (refer to Fig.2.3).

Based on the upon criteria, a set of twitter users in London were scanned and 65 of them were initially selected for implementing the methodologies of Chapter 3. For the data mining part, a script written in Python 2.7 is developed. Several libraries for authentication and data conversion are utilized (json, simplejson, oauth2, httplib2). In addition, Python Twitter, which is a Python wrapper around the Twitter API, is utilized.

The Python script is able to get the user Timeline information and return the most recently published 250 tweets. Each user's Timeline is stored later on in a csv file. The entire information of one generated tweet is stored (namely: "User's Name"; 'Latitude'; "Longitude"; "Place Name"; "Message ID"; "Source (i.e., smartphone, web, instagram etc.)"; "Timestamp (time, day, month, year)"; "Embedded Text (published micro-blogging content)"). Only the samples of active Twitter users who tweeted actively for at least 4 months and shared their geo-location information were considered.

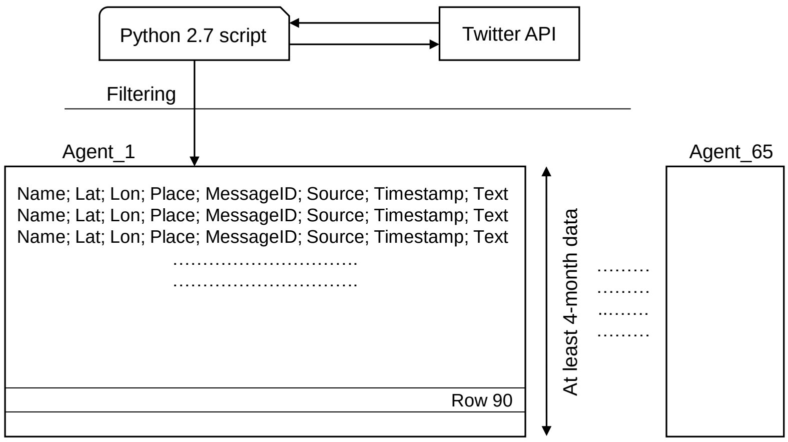


FIGURE 2.4.: Data Mining and Filtering

All 65 London are selected for analyzing individuals from only one region (i.e., city) which serves as a test zone. Focusing on the 65 twitter users who share geo-location information publicly, the followings are achieved: 1) by knowing the coordinates of a visited place from a user, the place is labeled and one knows the frequency and the exact times when a user returns at the same place; so one can categorize the place into home location, fixed, quasi-fixed and flexible activity; 2) the Point of Interest associated with the re-visited place can be retrieved with the use of maps (i.e., location coordinates reveal that it is an office, shopping center, university etc.) thus facilitating the validation of associating locations to activities; 3) the distance (either great-circle distance or traveled distance) between different places can be computed via using their geo-location.

However, limiting the search into users who share fully their geo-location information at all times introduces limitations to the sample collection. For instance, although Londoners use Twitter, it was observed that only a small fraction 3-5% had enabled the Tweet Location Feature on his/her Smartphone for sharing geo-location information while tweeting as described above. In future, the number of studied individuals can be increased via anticipating more Twitter users to activate the geo-location information sharing feature or by fusing user-generated data from smartcard logs and personalized navigators for increasing the sample representativeness. By working on a targeted sample of users with information-rich data though, one can draw general conclusions about their joint leisure activity participation and how it can be improved even if drawing general conclusions at the city level is not possible[2] .

2: Given the length of the 6400 posts that are collected from the 65 users over a 14-month time period, only a sample from the posts of one user is presented to provide an idea about the structure of the data.

After the data requirements of Chapter 3, in Chapter 4 methodologies for optimizing the location and time of joint leisure activities are developed. Those methodologies are based on optimizing the utility of activity participants; therefore, data sources from Chapter 3 can also be used in Chapter 4. Nevertheless, there is a significant difference that reduces the alternative data source options. Chapter 3 focused mainly on capturing individuals willingness for traveling certain distances to participate at different types of activities which requires individual-level geo-location data over a significant time period from the users in the study area. In Chapter 4, the optimization of location and time requires further information about the social networks of each user (the full list of friends and acquaintances of each user) because users should have some form of relationship to participate at joint leisure activities. For defining the social network structure of each individual and understanding with whom he/she has high chances to participate at common activities more qualitative information is needed. For example, smartcard logs provide information about the geo-location of a user and how he/she moves in the city with the use of public transport but they fail to provide any information about his/hers friends and acquaintances with whom he/she can participate in common activities. The same is true if we utilize solely the geo-location traces of users from their personalized navigators or their GPS-enabled smartphones since there is no information about potential friends with whom they participate at joint leisure activities. In contrary, Social Media have the advantage of providing richer qualitative data about the user because they hold information about the demographic characteristics of the user, the list of his friends and acquaintances and the level of closeness of the relationship of each user with his/her friends. In addition, oftentimes social media posts include tags of two or more persons indicating that at a specific geo-location a joint leisure activity took place.

Therefore, for optimizing the location and time of joint leisure activities with user-generated data the need of Social Media data sources is stronger that it was for simply capturing the utility satisfaction of users at participating at activities at different places and times of the day.

As described before, twitter was selected among the different social media options as a main data source because a number of users post geo-tagged tweets and there are no authentication requirements for retrieving those tweets with the use of the Twitter API. The Python Twitter, which is the Python wrapper around the Twitter API, is also utilized at chapter 4 for getting the user Timeline information and return the most recently published 250 tweets. Among all the retrieved information of users, raw data processing was performed in the same lines with chapter 3 with the additional requirement that the selected users have some kind of relationship (have exchanged tweets and are tagged together at similar locations). London is again selected as the study area and a more intense data crawling campaign resulted to the selection of 75 twitter users that satisfy the above-mentioned requirements[3] .

3: 10 more users compared to chapter 3

At this point, it should be noted again that the utilized raw data for implementing the developed methodologies in chapters 3, 4 do not cover all users in the study area (London) but a small fraction of them because of the very demanding data requirements. Nevertheless, there is no loss on generality due to that and the developed methodologies can be implemented in other study areas/cities in the exact same manner. For instance, if in future more twitter users share their geo-tagged locations publicly and tweet more actively in some study areas, then one can form bigger data samples with several more users and implement the same methodologies described in chapters 3, 4. This is very important because there is no need for generating a new set of methodologies for serving the specific needs of each study area.

After that, the data requirements in Chapter 5 are linked with the developed methodologies for adapting the departure times of public transportation services to the joint leisure activity demand. For this purpose, as described in Fig. 2.4, GTFS data of public transport services and their operational constraints are required together with user-generated geo-location data of all persons in the study area. For changing the departure times of public transport services without affecting the operational constraints, information regarding all those operational constraints should be provided (for instance, the mandatory resting times of drivers after each trip completion, the lowest and highest acceptable levels for the service frequency). In addition, GTFS data regarding the examined public transport operations should be provided for deriving the scheduled arrival times of the trips at each stop over the day and the topology of the network.

Given the above data requirements, in this thesis GTFS data from Sweden including the planned schedule of public transport modes

File Name	Size	File Name	Size
agency.txt	5KB	stop_-times.txt	242,000KB
calendar.txt	17KB	stops.txt	6,384KB
calendar_dates.txt	890KB	transfers.txt	1,020KB
routes.txt	218KB	trips.txt	11,628KB

Table 2.6.: GTFS data files and size (total size: 0.25GB)

for the period 13 February 2016 - 17 June 2016 is utilized. The data includes the files of Table 2.6 and also information about the lowest and highest starting headway limits between successive trips. Working on the same study area in all chapters provides more homogeneity, but:

- the diverse and demanding data requirements of each chapter
- the nature of the methodologies of this thesis which can be applied at different study areas without loss of generality

led to the selection of another study area (Stockholm, Sweden instead of London, UK) for implementing the methodologies of chapter 5. The reason behind the selection of Stockholm is the publicly available API for downloading files in the standard GTFS format of all planned public transportation trips throughout Sweden[4] .

4: It can be accessed at https://www.trafiklab.se/api/gtfs-sverige-2.

To derive the planned schedules of public transport modes, a library was developed in Python 2.7. The library processes .txt files and converts/stores them to an SQL database. This facilitates data queries and enables web-based visualization of the public transport operations with the use of OpenStreetMap (via OpenLayers, an open-source JavaScript library[5]). The developed Python GTFS library: i) converts .txt files to sql database tables, ii) can query public transport routes from the database tables, iii) creates new files containing the planned trips for each route in ascending order (starting from the earlier morning trip to the latest night trip).

5: http://www.openlayers.org/api/OpenLayers.js

After applying the Python GTFS library, the planned trips for every public transportation service are sorted and for each trip one can get the planned arrival time at every station. In particular, the thesis focuses on two key central bus lines in Stockholm which are bidirectional (lines 1 and 4) and can be seen as 4 independent services because every line direction has another structure and another set of constraints. Those are bus service 1 (bus line 1, direction 1 (Essingetorget to Stockholm Frihamnen)), bus service 2 (bus line 1, direction 2 (Stockholm Frihamnen to Essingetorget)), service 3 (bus line 4, direction 1 (Gullmarsplan to Radiohuset)) and service 4 (bus line 4, direction 2 (Radiohuset to Gullmarsplan)). In Fig.2.5 the stops of the central bus lines in Stockholm are presented for reference purposes.

Apart from public transportation data, the data of travelers that are heading to joint leisure activities is also retrieved with a twitter crawling campaign in the study area of Stockholm. For those users, the geo-tagged locations of their tweets should also be known and, finally, only the geo-location information from 62 individuals from Stockholm is selected. Those 62 individuals are selected according to the following

FIGURE 2.5.: Central Bus lines in Stockholm - Network Representation of bus lines (source: `http://sl.se/ficktid/karta/vinter/SthlmInnerstad.pdf`)

criteria: i) they change geo-location outside working hours (from 15:30-19:30); ii) the location they are heading is not related to work/home, but it is a leisure activity location for them; iii) their travel origin is within walking distance from at least one bus station of lines 1 and 4; iv) they arrive at the location of the joint leisure activity within a time variance of less than 25min.; therefore, it is assumed that they participate in the same activity. Criterion (iv) is the strongest assumption since multiple joint leisure activities can occur nearby concurrently. However, by depending solely on Twitter data one cannot justify the potential splitting of a joint leisure activity into multiple concurrent ones with a high degree of confidence. The assumption (iv) would have been avoided if users were always tagged together at the same post at the joint leisure activity location, but such tags are not very common. For this reason, revealed-preference surveys could have been used for those 62 individuals in order to elicit information regarding the number of the common activity participants and avoid assumption (iv). However, one should be aware that revealed preference surveys are labor intensive and cannot scale up for covering the entire study area population.

Finally, the data requirements of chapter 6 are related with the multimodal journey planning of individuals who are willing to travel between two points for participating at a joint leisure activity. In this chapter each transport mode is modeled via a layer-based system and the entire study area transport network is represented with the integration of different layers of a GIS system. For this reason, the topology (i.e., location of intermediate stops) and schedules of different transport modes should be provided for the development of such system. Based on that, the transportation network is modeled as a set of nodes (stops) which are connected with links. One important characteristic is that one might be able to travel from one node to another with the use of different modes (walk, drive, use public transport etc.). Therefore, there is not only one link that connects two nodes because there might be multiple transport mode options that connect the same nodes and each transport mode connection is represented by a different link. The main data requirement for implementing this layer-based system that utilizing nodes and link connections is the additional travel time information of each link for different times of the day and the fuel consumption while traversing that line. In addition, for considering also the user's preferences on selecting the optimal path, each user should also provide data regarding:

- the maximum distance that the he/she is willing to walk
- the maximum distance that the he/she is willing to travel
- the maximum transfers between different transport modes that he/she is willing to perform
- any potential transport modes that he/she is not willing to use

Therefore, there is a broad set of information regarding the (a) transport network of the study area (the travel times and the fuel consumption while traversing each link) and (b) the users' preferences such as the maximum walking distance, the maximum number of transfers and the exclusion of transport modes. The challenge behind solving the joint leisure activity multi-modal trip optimization subject to the set of traveler preferences is computational as described in chapter 6 given the the vast set of alternative links in a study area that can be combined for performing a trip. As a result, the methodologies of chapter 6 are focused on this direction and are general enough for enabling their application to different study areas (i.e., different cities) with the only requirement to modify the data input accordingly.

Given the large and heterogeneous data requirements of chapter 6 from both the transport network's topology/operations and the travelers' preferences, several validation tests are implemented on randomized networks that contained $100, 200, 300, 400, 500, ..., 3000$ nodes-stations. In order to minimize the effects of the optimal multi-modal journey selection due to the networks' topology, both sparse and dense networks were examined. The execution environment of each randomized network test is described by its number of nodes, the number of links, the level of density, the number of transportation modes and the number of constraints (Table 2.7 summarizes the details of the utilized networks in this thesis). Each link connects an origin with a destination node via a distinct transport mode and more links than one may connect the same origin-destination pair. The traveler preferences are also set with the use of randomized maximum values (constraints) regarding the total travel's fuel consumption, covered distance, walking time and number of transfers. Those maximum values that represent the users' preferences can be simply updated in practical applications with the use of individual-level data from revealed preference surveys.

Table 2.7.: Density of Randomized networks and Users' Preferences (represented by Constraint Values)

Randomised Network					
Network Nodes	Purpose-Built Links	Ratio of Nodes/Links	Number of Modes	Number of Constraints	Limits of Constraints (units)
100	1000	10.00%	2	4	15.6/135/19/1
200	2009	9.95%	3	4	13.6/115/17/2
1000	11000	9.25%	3	4	15.4/122.9/15/5
1400	14000	10.57%	4	4	22.5/186/13/4
1600	16000	10.38%	4	4	31/285/25/2
1800	18000	10.18%	5	4	66/470/34/3
2600	26000	10.41%	5	4	200/1550/18/2
500	26000	1.98%	3	4	10.7/91/11/3
300	20400	1.47%	3	4	2/12.4/21/2
400	24000	1.67%	4	4	6/39.6/23/3
2800	28000	10.25%	6	4	89/540/14/4
3000	30000	10.08%	6	4	161/1040/28/3

Capturing users' willingness to travel certain distances

3.

Selecting a location and time for joint activity participation among multiple individuals (agents) in a dense city is not a trivial task due to the lack of knowledge regarding agents' preferences and the presence of numerous Places of Interest (POIs). Nowadays, new data sources unveil new opportunities through the collection and analysis of more detailed, user-generated data that can be utilized for improving the level of information regarding agents' profiles and preferences.

As discussed at the background theory section, publicly available user-generated data has the potential to improve the understanding of travelers' patterns and mobility preferences including transport mode selection, their departure and arrival times, frequency and scope of undertaken trips. Mining, handling and analyzing such data (i.e., Social Media data (SM), Floating Car data (FCD), Mobile Phone data (MP) and Smart Card data (SC)) for gaining insights into transportation dynamics is a challenging task due to the data volume and level of dynamism; hence, the development of advanced data-analytic techniques is required.

Regardless the posed challenges, user-generated data can be utilized to improve our understanding on human mobility and develop individual-level models for capturing the mobility and activity habits of individuals.

Building on top of the basic assumption of Activity-based models (refer to [5] , [6], [7], [8]): "individuals' trips are means to participate in activities and are not undertaken without purpose due to their inherent dis-utility", the potential of utilizing more detailed, user-generated data for understanding users' willingness to travel in order to participate in different sets of activities during different day times and day types (i.e., weekdays/weekend) is examined.

[5]: Axhausen et al. (1992), 'Activity-based approaches to travel analysis'

In a first stage, users' data logs are examined and their mobility and activity patterns are retrieved automatically after a learning phase. Later on, a utility-maximization model is developed for capturing users' willingness to travel a certain distance for participating in different types of activities via an individualistic, utility-maximization model. Finally, users with commonalities are clustered based on their willingness to travel similar distances for participating in certain types of activities.

In the literature, [69] utilized SM data for capturing the demand variation during special events and [9] utilized MP data for correlating the mobility patterns of an individual with the mobility patterns of his friends and acquaintances. In the study of [28] and [29] Bluetooth devices were distributed to people to collect mobility data and study the characteristics of co-location patterns among people. In addition, [14], [15], [3], [18], [19], [17] and [70] , utilized MP data for predicting

[69]: Pereira et al. (2014l), 'Using data from the web to predict public transport arrivals under special events scenarios'

[70]: Shang et al. (2012), 'Predicting dynamic transportation demand with mobility data'

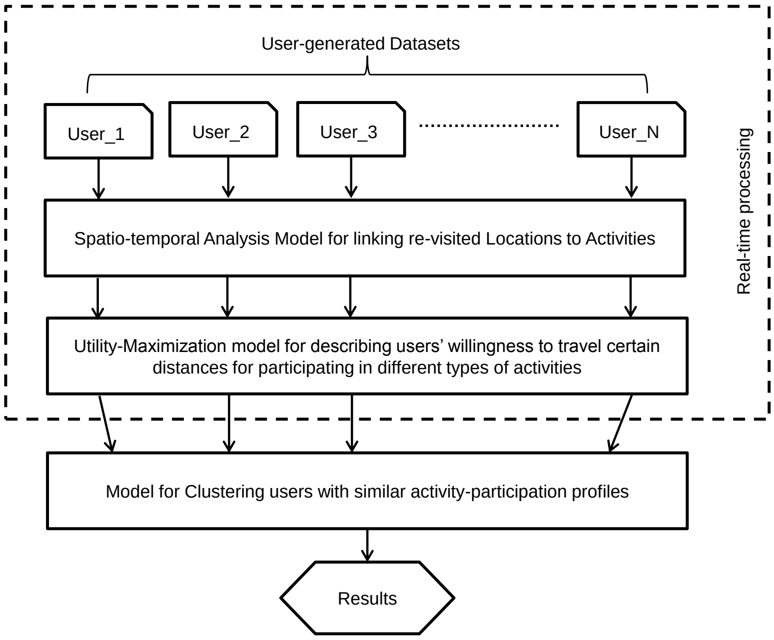


FIGURE 3.1.: Work Structure

the mobility patterns of individuals. However, little has been done in the area of activity-participation analysis which can facilitate the development of new applications for suggesting common activities to users with social ties based on their willingness to participate simultaneously in similar activities being held in nearby locations.

In more detail, there is no systematic way for utilizing BigData for capturing and later on modeling individuals' willingness to travel certain distances for participating in activities over different time periods. At first, an individual-level analysis on the potential of user-generated data towards that direction is required and this work contributes on that via introducing models for unfolding individuals' decisions based on pattern recognition and models for estimating the distances individuals' are willing to travel for participating in certain types of activities via utility-maximization techniques which facilitate the identification of similarities among travelers. Another contribution is that apart from analyzing HBW trips which are usually captured by household surveys, new activity types can be investigated since user-generated data offers continuous feeds that can be utilized for analyzing more detailed trip purposes.

In view of the above, a set of techniques for processing user-generated data in order to capture the willingness of users to participate in certain types of activities are introduced (refer to Fig.3.1).

3.1. Daily Pattern Recognition Model

Forecasting Individual's Mobility based on Pattern Observation

Continuous updated, user-generated data is utilized to capture less frequent trips and improve the understanding of individuals' mobility behavior. Revealed preference surveys are typical methods for capturing individual's mobility patterns over time. Nevertheless, the information provided from surveys is static - even if the form of the survey is detailed including questions about mobility variations from day to day. Although static information is adequate for capturing individuals' main trips (i.e., trip to and from work), it fails to capture the change on users' behavior over time and the participation to non-recurrent activities.

In general, user-generated data includes geo-tagged locations and the time stamp of users' interactions. In this section, a Pattern Recognition Model (PRM) is proposed with a dual scope: a) to process automatically mass volumes of user-generated data and capture users' mobility patterns; b) to link activity types to geo-tagged locations based on spatio-temporal analysis of users' interactions. The PRM utilizes data from one individual's interactions over a significant time period and retrieves his/her preferences. New data feeds are also handled automatically by the PRM yielding an improvement of capturing users' patterns.

The PRM of one individual is a model that describes the mobility patterns of an individual in a probabilistic form, considering also the activity dimension and not only the location information. The values of the parameters of one individual's PRM are defined after a learning phase by utilizing user's data logs. The spatio-temporal mobility and activity patterns of a user during different day types (week-weekend days) are defined after a learning phase:

$$P(I,k,t,L_m,A_n) = \frac{N(I,k,t,L_m,A_n)}{\sum_{A_n \in A} \sum_{L_m \in \Lambda} N(I,k,t,L_m,A_n)} \tag{3.1}$$

, where $N(I,k,t,L_m,A_n)$ is the number of times that user I was in location L_m and participating in an activity A_m at time t and day type k.

Matrix P has $[I \times k \times t \times L_m \times A_n]$ elements showing which is the probability of a user to be in a location and participate in an activity for different times and different day types. Those probabilities are calculated after accumulating the mobility footprint of a user *over time* derived from his/her SM interactions.

Over time refers to the aggregation of time instances at which the user generated social media content. This is not a static time period and its duration differs from individual to individual. This thesis studied user-generated data which was consistently generated over more than one year[7] forming a time period of 14 months. Nevertheless, a portion of

7: November 2012 – January 2014

individuals remained inactive for some time periods and their mobility patterns were constructed from a condensed dataset. To that point, it should be mentioned that two threshold values were introduced to ensure the quality of data when retrieving individual's mobility footprint. The first threshold was the time period of interactions (all individuals who were active for less than 4 months were excluded). Hence, the lower bound was set to a 4-month period of interactions. The second criterion was the volume of data. In that case, each individual should have more than ninety (90) interactions at which he/she shared also his/her geo-location; otherwise, the individual was excluded from the sample.

Having calculated one individual's probability of being at a particular location and participating in a particular activity for a time t and day type k, the daily activity plan of the individual can be generated via deterministic modeling by using single point estimates, where the output is a sequence of states (refer to Fig.3.2).

Nevertheless, even if there is a high probability for a user, I, to be present in a location L_m at time t, his/her previous location should also be taken into consideration without resulting though to a pure no-memory process (Markov Chain). If the user was at location L_{m-1} during the previous simulation time step and there is no observed travel history from location L_{m-1} to L_m, then the user cannot be assigned to location L_m at time t. To model that effect, the transition probability over two consecutive time instances is introduced:

$$a_{i,j} = P(S_{t+1} = q_j | S_t = q_i) \tag{3.2}$$

where the transition probability returns the probability to transfer from state q_i to state q_j.

Let assume that the individual is at location L_m and participated in activity A_n at time t. At time $t+1$ the individual is highly probable to be at location L_{m+1} and participate in activity A_{n+1} according to the probability matrix; hence, this option might be the current single point estimate. Nevertheless, if the transition probability for transferring from L_m, A_n to L_{m+1}, A_{n+1} at time t is zero such transfer is not allowed and the next candidate is selected. Therefore, the transition probability is utilized as a check point for re-assuring that the proposed transition is feasible.

Linking Activity types to visited Locations

Deriving the activity type performed in a particular location by a user is another issue. Based on one individual's interactions on different locations and at different day times and day types, users' mobility patterns can be identified. For instance, Fig.3.3 shows one user's interactions over time (x-axis), space (y-axis) and level of re-currency (z-axis), where F_i is the distance of a location from the home of the examined user in km.

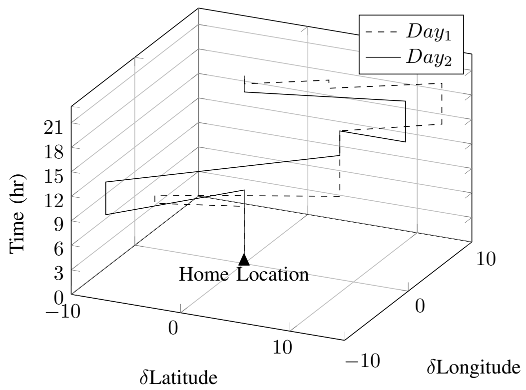


FIGURE 3.2.: Simulating the Mobility Trajectories of a Random Individual over two Weekdays

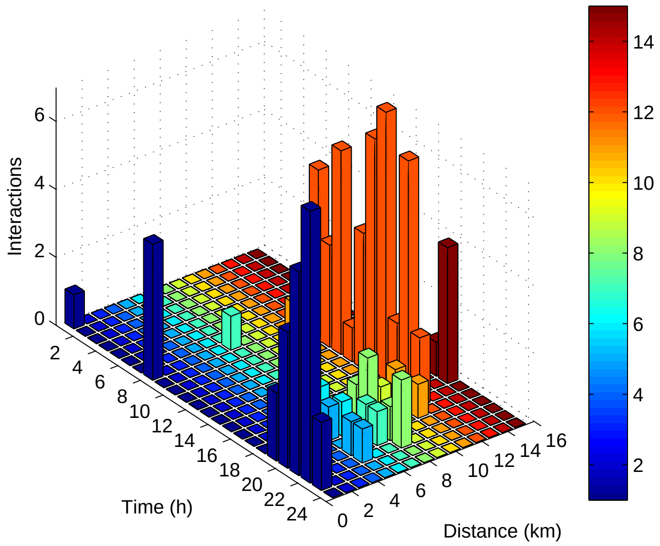


FIGURE 3.3.: Vizualization of individuals' mobility patterns over time, where Distance (km) is the distance from the Home location

Individuals depart and arrive at locations to participate in activities. However, the same location can be the home for one user, the work place for another and the location of a flexible activity for a third. Due to that, users' interactions over a significant time period are analyzed and for each user, I, every visited location, L_n, is assigned to an activity type:

$$A(I, L_n) = A_m \tag{3.3}$$

, where

$$A_m = \begin{cases} A_1 : \text{Home} \\ A_2 : \text{Fixed Activity} \\ A_3 : \text{Quasi-Fixed Activity} \\ A_4 : \text{Flexible Activity} \end{cases}$$

, I is the ID of the examined individual and $L_n \in \Lambda$ is the identification number of the location.

Fixed activities are recurrent activities such as work, school, university, language classes, gym, etc. Quasi-fixed activities have a recurrent nature with lower frequency (i.e., the location of a favorite restaurant, park etc.) while flexible activities are linked to non-habitual, recreation activities.

In order to associate a location with the "home activity" of a user, $A(I, L_m) = A_1$, let:

$$x^{t,k} = \begin{cases} 1 & \text{if an individual generates data at time } t \text{ and day } k \\ 0 & \text{if there is no interaction} \end{cases}$$

Let also $Y = \{y_1^{t,k}, y_2^{t,k}, ..., y_N^{t,k}\}$, where

$$y_i^{t,k} = \begin{cases} 1 & \text{if an individual is at location } L_i \text{ at time } t \text{ and day } k \\ 0 & \text{in all other cases} \end{cases}$$

Moreover, let

$$w^{t,k} = \begin{cases} 1 & \text{if } k \text{ is a weekday} \\ 0 & \text{if } k \text{ is weekend day} \end{cases}$$

If someone examines the temporal variations of SM users' interactions, he/she will observe that there is a time period at which users are inactive. After analyzing users' data from many weekdays, the time period of in-activeness for each user can be estimated. During that period, we assume that the user is located at home and performs an activity that does not require action (i.e., relaxing/sleeping).

A lower bound Max_T is set as the latest time at which an individual generated content and is derived after examining his/her interactions over time. Furthermore, Min_T is set as the upper bound and it is the earliest time in the morning at which an interaction is observed.

For deriving the location linked to home activity automatically, it is assumed that individuals interact more from locations close to their home early in the morning and late at night during weekdays. Therefore, if location L_i is assigned to the home activity of a user, then it should satisfy the following inequality:

$$\begin{aligned} &\sum_{k=D_1}^{k=D_M} \sum_{t=Min_T}^{t=Min_T+2hr} y_i^{t,k} w^{t,k} + \sum_{k=D_1}^{k=D_M} \sum_{t=Max_T-3hr}^{t=Max_T} y_i^{t,k} w^{t,k} \geq \\ &\sum_{k=D_1}^{k=D_M} \sum_{t=Min_T}^{t=Min_T+2hr} y_l^{t,k} w^{t,k} + \sum_{k=D_1}^{k=D_M} \sum_{t=Max_T-3hr}^{t=Max_T} y_l^{t,k} w^{t,k} \quad \forall L_l \in \Lambda \end{aligned} \tag{3.4}$$

This rule is derived after applying computational learning to the users' datasets and is modifiable based on the observed data of citizens when studying different cities/areas. In this work, the rule was generated after studying Social Media data from 65 users in London collected from November 2012-January 2014 containing 6,400 interactions. This rule had an accuracy of 92% on assigning users' home activity to the correct location. Results are presented in Fig.3.4 where the y-axis

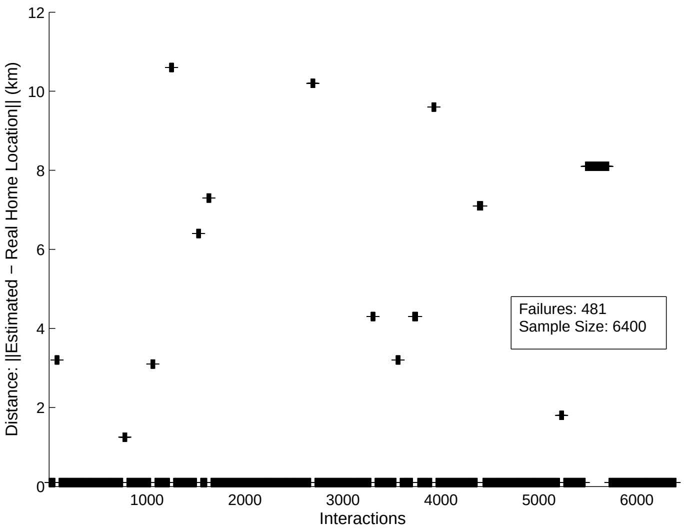


FIGURE 3.4.: Home Activity Estimation Errors

shows the distance between the estimated and the real home location in km.

To calculate the distance of other locations from the locations linked to the home activity, the Haversine formula is utilized. Therefore, the distance between the location L_i of a location and the Home location, L_h, is set as F_i (in km) after its computation with the Haversine formula.

The Harversine formula is utilized in order to introduce a method for distance calculation and comparison of distances willing to be traveled by users to participate in different activities. This work proposes also a simple categorization into four generic activity groups in order to link locations with activities while maintaining a high level of accuracy. Fixed-activity centroids are recurrent locations that satisfy the following criterion:

$$\sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} y_i^{t,k} \geq 0.10 \sum_{k=D_1}^{k=D_M} \sum_{t=t_0}^{t=t_N} x^{t,k} \tag{3.5}$$

Similarly, quasi-fixed activity centroids are locations that accumulate less than 10% but more than 5% of the total number of the user's interactions and flexible activity centroids those accumulating less than 5%.

3.2. Utility Maximization Model for capturing users' willingness to travel a certain distance in order to participate in different activities

Individuals select to participate in different types of activities at different times of day based on a decision-making mechanism that seeks to maximize their level of satisfaction. In an attempt to model the decision-maker's choice of traveling a certain distance for participating in a certain type of activity, a utility function is defined. To decompose the activity selection model, a time discretization scheme is adopted where each time step lasts for one hour. Time is indexed by $t = \{1, ..., T\}$ and the choice options are indexed by $j = \{1, ..., J\}$ where F_j is the traveled distance between two consecutive activities.

$$j = \begin{cases} 1 : F_j \leq 250m \\ 2 : 250m < F_j \leq 500m \\ 3 : 500m < F_j \leq 750m \\ 4 : 750m < F_j \leq 1km \\ 5 : 1km < F_j \leq 1.5km \\ 6 : 1.5km < F_j \leq 3km \\ 7 : 3km < F_j \leq 5km \\ 8 : 5km < F_j \leq 10km \\ 9 : F_j > 10km \end{cases}$$

For each day type, an index of satisfaction for participating in different activity types with respect to their distance from the previous location is defined in the form of a linear utility function:

$$V_{tj}(k) = \alpha_j(k) + \beta_j(k)A_t(k) \tag{3.6}$$

In the above notation, $A_t(k)$ varies across different times of the day and represents the activity type (home, fixed, quasi-fixed or flexible) in the form of a categorical variable. In addition, α is a scalar utility term representing individual's preference for alternative j.

For modeling the qualitative attribute, $A_t(k)$, a $0-1$ attribute scheme is introduced with coding convention as presented below.

Table 3.1.: Density of Randomized networks and Users' Preferences (represented by Constraint Values)

	z_α	z_β	z_γ
Home	0	0	0
Fixed	1	0	0
Quasi-fixed	0	1	0
Flexible	0	0	1

and Eq.3.6 results in:

$$V_{tj}(k) = \alpha_j(k) + \beta_{j,1}(k)z_{\alpha,t}(k) + \beta_{j,2}(k)z_{\beta,t}(k) + \beta_{j,3}(k)z_{\gamma,t}(k) \tag{3.7}$$

The random utility of alternative i, $U_i(t,k)$, for an individual can be described by a random utility model:

$$U_{tj}(k) = V_{tj}(k) + \epsilon_{tj}(k) \tag{3.8}$$

where $\epsilon_{tj}(k)$ is the unobserved component of the utility function and can be represented as a random variable since it includes the impact of all the unobserved variables which influence the utility of selecting a specific alternative.

With the assumptions that errors follow a Gumbel distribution, are independent and identically distributed, the probability of selecting an alternative $P_{ti}(k)$ can be expressed via a multinomial logit model:

$$\begin{aligned} P_{ti}(k) &= P\Big(V_{ti}(k) + \epsilon_{ti}(k) \geqslant max_{j\in\{1,\ldots,J\}} V_{jt}(k) + \epsilon_{tj}(k)\Big) \\ &= \frac{e^{V_{ti}(k)}}{\sum_{j=1}^{j} e^{V_{tj}(k)}} \end{aligned} \tag{3.9}$$

Other assumptions, such as Thurston assumption which leads to a probit model can be utilized. In our case though, it is assumed that errors follow a Gumbel distribution yielding a multionomial logit model which has been also used extensively in problems related to transport mode choice. In more detail, it is assumed that alternative options are different to users and that a newly introduced alternative expanding the choice set will not change the preference of a prior alternative towards other prior alternatives. In other words, the relative probabilities of selecting a quasi-fixed or a fixed activity do not change if a flexible activity is added as an additional possibility. This case might not hold if the individual pre-plans not one, but several inter-correlated activities at once; however, in this analysis we are not considering that instance since in each time index is selected one activity.

The parameters to be estimated are the terms $\alpha_j(k)$, $\beta_j(k)$ for each individual. Since the entire set of users' interactions is not available, the parameters can be estimated by maximizing the likelihood function. To estimate the MNL model, the maximum-likelihood estimator is required and it is the same regardless of whether one maximizes the likelihood or the log-likelihood function, since log is a strictly monotonically increasing function. The log-likelihood of user's activity-mobility observations at time t and day type k is:

$$\begin{aligned} \ell\big(\alpha_j(k), \beta_j(k)\big) &= \sum_{t=1}^{T}\sum_{j=1}^{J} y_{tj} ln\big(P_{tj}(k)\big) \\ &= \sum_{t=1}^{T}\sum_{j=1}^{J} y_{tj}\big(V_{tj}(k) - ln\sum_{j=1}^{J} e^{V_{tj}(k)}\big) \end{aligned} \tag{3.10}$$

where $y_{tj} = 1$ if the individual chose alternative j at day type k and 0 otherwise.

The observed values of the traveled distance by the examined user are computed by calculating the distance between user's location at time instant t and day type k and the observed location at time

$t-1 \in t = \{1, ..., T\}$ and at the same day using the Haversine formula. Then, parameters $\alpha_j(k), \beta_j(k)$ are estimated as the values that maximize the log-likelihood function:

$$\max_{\alpha_j(k),\beta_j(k)} \ell\big(\alpha_j(k), \beta_j(k)\big) \tag{3.11}$$

[71]: Berndt et al. (1974), 'Estimation and inference in nonlinear structural models'

resulting to a non-linear optimization problem for which the optimization algorithm BHHH proposed by [71] is applied. BHHH estimates coefficients through optimization when a non-linear model is fitted to data. The coefficient values are updated in an iterative approach beginning with a starting set of values and iterations continue until convergence.

3.3. Clustering Users based on their Utility Model Similarities

After deriving users' patterns and their utility models, user profiles are clustered based on their willingness to travel similar distances to participate in certain activities.

In the clustering phase, users' are treated as entities with no personal information via utilizing a randomized ID as the only identification label for each user.

In order to compute the distance between two users, we compare the distance between their utility-maximization models over time. The distance represents the similarity of users' willingness to travel similar distances for participating in similar activities.

For instance, the probability of traveling a distance (0.25km, 0.5km] for participating in a quasi-fixed activity at time t and day type k is calculated based on the utility-maximization model of each user. Then, the probability is compared with the same probability calculated for another user and the observed divergence is the distance between those two users (user-couple). Comparing the probabilities of a couple of users for traveling a certain distance range for participating in a similar activity over time $t = \{1, ..., T\}$ and different day types, k, returns the probability distance between that couple:

$$CL_{I_1,I_2} = \frac{1}{|A_m|T} \sum_{k=0}^{k=1} \sum_{t=0}^{T} \sum_{j=1}^{J} \sum_{A_m=0}^{|A_m|} \left| \frac{e^{V_{I_1,tj}(k)}}{\sum_{r=1}^{J} e^{V_{I_1,tr}(k)}} - \frac{e^{V_{I_2,tj}(k)}}{\sum_{r=1}^{J} e^{V_{I_2,tr}(k)}} \right|^2 \tag{3.12}$$

, where I_1 and I_2 are the pseudo-IDs of the compared users, k the day type (in our study we consider two day types: week-weekend), t the time and $T = 23$ if an one-hour time discretization scheme is applied, A_m the activity type and:

$$V_{tj}(k) = \begin{cases} \alpha_j(k) : \text{if } A_m = \text{Home} \\ \alpha_j(k) + \beta_{j,1}(k) z_{\alpha,t}(k) : \text{if } A_m = \text{Fixed} \\ \alpha_j(k) + \beta_{j,2}(k) z_{\beta,t}(k) : \text{if } A_m = \text{Quasi-Fixed} \\ \alpha_j(k) + \beta_{j,3}(k) z_{\gamma,t}(k) : \text{if } A_m = \text{Flexible} \end{cases}$$

The distance CL_{I_1,I_2} can be computed for all user couples $I_1, I_2 \in I$ resulting to a $N \times N$ square matrix:

$$[CL] = \begin{pmatrix} CL_{0,0} & CL_{0,1} & \cdots & CL_{0,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ CL_{j,0} & CL_{j,1} & \cdots & CL_{j,N-1} \\ \vdots & \vdots & \ddots & \vdots \\ CL_{N-1,0} & CL_{N-1,1} & \cdots & CL_{N-1,N-1} \end{pmatrix} \tag{3.13}$$

On a second stage, the Density-based Spatial Clustering of Applications with Noise (DBSCAN) proposed by [72] for users' clustering based on probability distance is utilized (refer to Alg.1).

[72]: Ester et al. (1996), 'A density-based algorithm for discovering clusters in large spatial databases with noise'

Algorithm 1: Modified DBSCAN for users' Clustering based on CL distance

Set a threshold CL distance value, τ below which users are assigned to the same cluster;
Set a counter i=0;
Initialize an array $Q = \varnothing$ for tracking users' IDs;
for *each user* $u \notin Q$ **do**
  $Q = Q \cup \{u\}$;
  Set $i = i + 1$;
  Initialize a new Cluster $C_i = \varnothing$ and set $C_i = C_i \cup \{u\}$;
  for *every user* $v | CL_{uv} < \tau$ **do**
    $C_i = C_i \cup \{v\}$;
    $Q = Q \cup \{v\}$;
  end
end
In the second stage of DBSCAN implementation, clusters that include similar users are merged;
while $C_i \cap C_j = \varnothing$ *for all* i, j **do**
  if *user* $v \in C_i$ *and* $v \in C_j$ **then**
    Merge clusters: $C_i = C_i \cup C_j$;
  end
end

The results of clustering depend on the sequence of the selection of SM users, $u \in Q$, in Alg.1. The optimum set of clusters can be selected automatically via setting a threshold value $m \in \mathbb{Z}_+$ which terminates the implementation of Alg.1 if more than m clusters are generated.

3.4. Retrieving users' willingness to travel for participating in different activities

The model system for retrieving automatically users' willingness to travel certain distances for participating in different activity types over different days and times is tested for 65 Social Media users residing in London after processing their publicly available, user-generated data including geo-tagged locations and time stamps of interactions. The dataset contains a sum of 6,400 interactions collected through a crawling campaign from November 2012-January 2014 and the gender and age categories of the sampled SM users is presented in Fig.3.6.

London was selected for analyzing individuals from only one region (i.e., city) which will serve as a test zone. In future, one can test also whether user-generated data, generated at the same region from external sources (i.e., Smart Card logs) can be fused with Social Media data logs to increase the sample size of the analysis. Nowadays, the open-data policy of Transport for London offers the collected Oyster Card data logs to the public, but they miss the characteristics of user-generated data feeds since they are pre-processed to ensure anonymization. Oyster Card data logs are stored as scattered trips among public transport stops without linking them to the passengers who undertook those trips. This hinders the analysis of travelers' daily patterns since there is no information regarding trips' ownership.

By restricting the research area to London, the main focus is placed on Social Media users who share geo-location information publicly because this was vital for the scope of our activity for the following reasons: 1) by knowing the coordinates of a visited place, the place is labeled and one knows the frequency and the exact times when a user returns at the same place; so one can categorize the place into home, fixed, quasi-fixed and flexible activity; 2) the Point of Interest associated with the re-visited place can be retrieved with the use of maps (i.e., location coordinates reveal that it is an office, shopping center, university etc.) thus facilitating the validation of associating locations to activities; 3) great-circle distance among different places can be computed via using their geo-location.

However, limiting the search into users who share fully their geo-location information at all times introduces limitations to the sample collection. For instance, although Londoners use Twitter, it was observed that only a small fraction 3-5% had enabled the Tweet Location Feature on his/her Smartphone for sharing geo-location information while tweeting. The Tweet Location Feature on Smartphones is off by default, and the majority of Twitter users are not aware of its existence at all or they refuse to activate it for mitigating data privacy risks. Due to that, the main focus was on active Twitter users from London who had a GPS-enabled Smartphone and had the Tweet Location Feature activated when tweeting. On a second step, only users who tweeted actively over a significant time period[8] more than ninety (90) times were considered in order to obtain enough historical data for pattern recognition. Since the research scope was focused on understanding

8: at least 4 months

mobility at the individual level, only healthy samples of information-rich user-generated data were gathered instead of crawling Social Media blindly and storing information with little relevance to the topic. In future, the number of studied individuals can be increased even by anticipating more Twitter users to activate the geo-location information sharing feature or by fusing user-generated data from Oyster Card logs with Twitter data for increasing the sample representativeness. Hence, by working on a targeted sample of agents with information-rich data one can draw general conclusions about individuals and their pattern similarities towards activity participation while avoiding drawing general conclusions at a higher level (i.e., insights on mobility patterns in London).

For the data mining part, a script written in Python 2.7 was developed. Several libraries for authentication and data conversion were utilized[9] . In addition, Python Twitter, which is a Python wrapper around the Twitter API, was utilized.

9: json, simplejson, oauth2, httplib2

The Python script was able to get the user Timeline information and return the most recently published 250 tweets. Each user's Timeline was stored later on in a csv file. The entire information of one generated tweet is stored (namely: "User's Name"; 'Latitude'; "Longitude"; "Place Name"; "Message ID"; "Source (i.e., smartphone, web, instagram etc.)"; "Timestamp (time, day, month, year)"; "Embedded Text (published micro-blogging content)"). As discussed before, only the samples of active Twitter users who tweeted actively for at least 4 months and always shared their geo-location information were considered. Therefore, the collected data has the form of Fig.3.5.

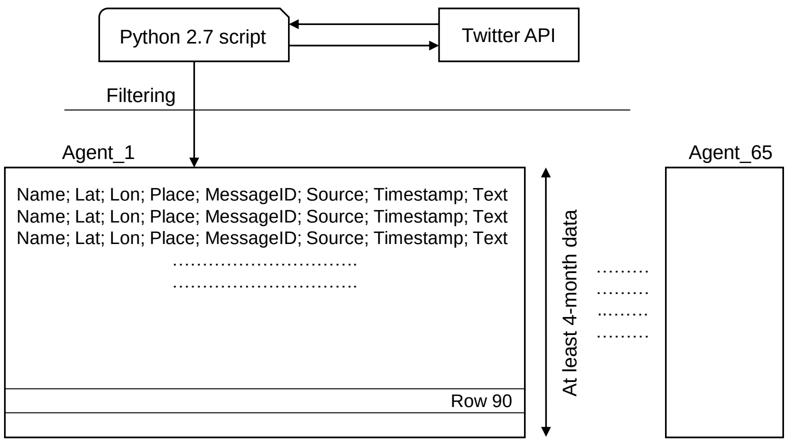


FIGURE 3.5.: Data Mining and Filtering

For each social-media user, a PRM model capturing his mobility/activity patterns and a Utility-maximization Model is developed showing the distance he/she is willing to travel to participate in different types of activity during the day based on the methods described in Section 3.1.

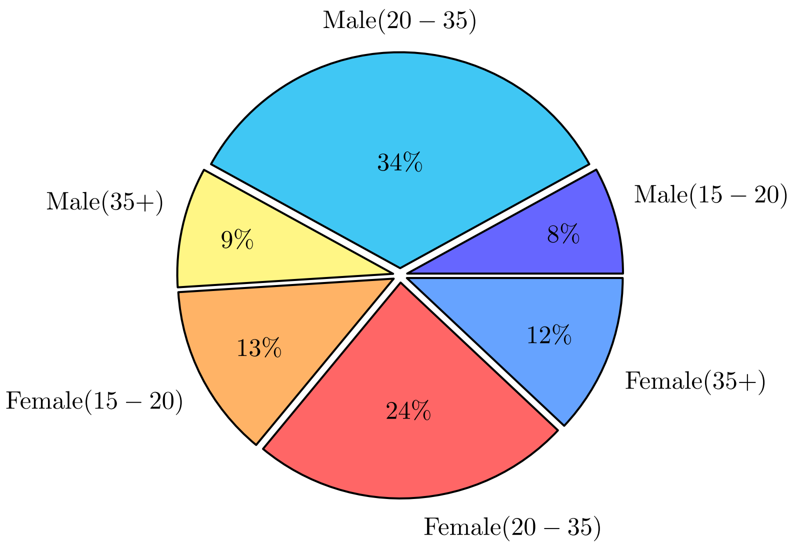


FIGURE 3.6.: Gender and Age Category of Social Media Users

The results of the automatic generated Utility-Maximization model of one individual (in this case, Social Media user #31) are presented in Fig.3.7, 3.8. For each time of day $t = \{0, ..., 23\}$ the distance willing to be traveled by the user for participating in an activity (home, fixed, quasi-fixed, flexible) is represented with an arrow. From Fig.3.7, 3.8 one can observe that user #31 is willing to travel more to reach home or participate in a flexible activity at late night times. In addition, he/she is willing to travel more km to participate in fixed activities (i.e., work) during early morning hours (7-9am). Similar figures can be derived for all users showing their willingness to travel certain distances for participating in different activity types during the day as they were calculated from their utility-maximization models.

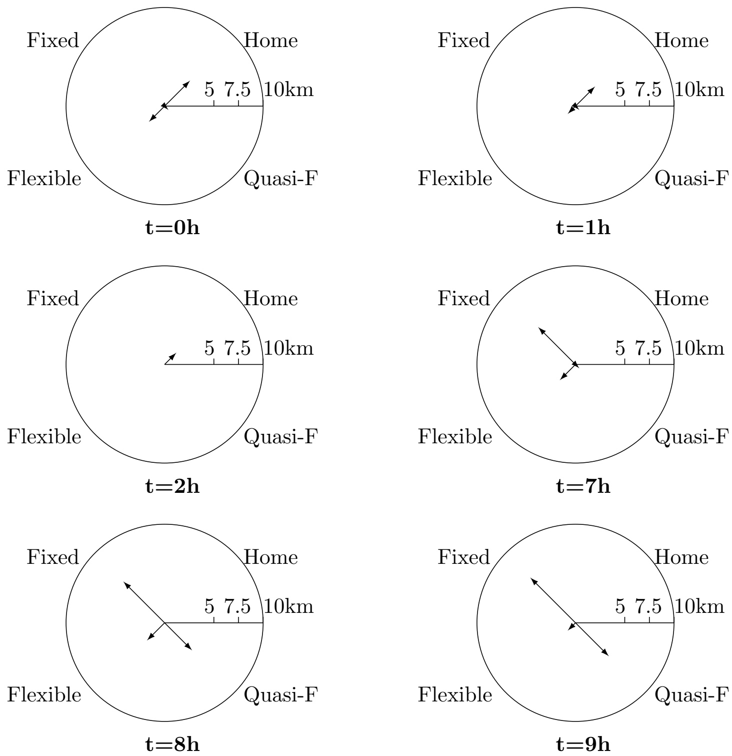


FIGURE 3.7.: 2D representation of the distance willing to be traveled by SM user #31 for participating at a specific activity type at late night and early morning hours retrieved from his Utility-Maximization model

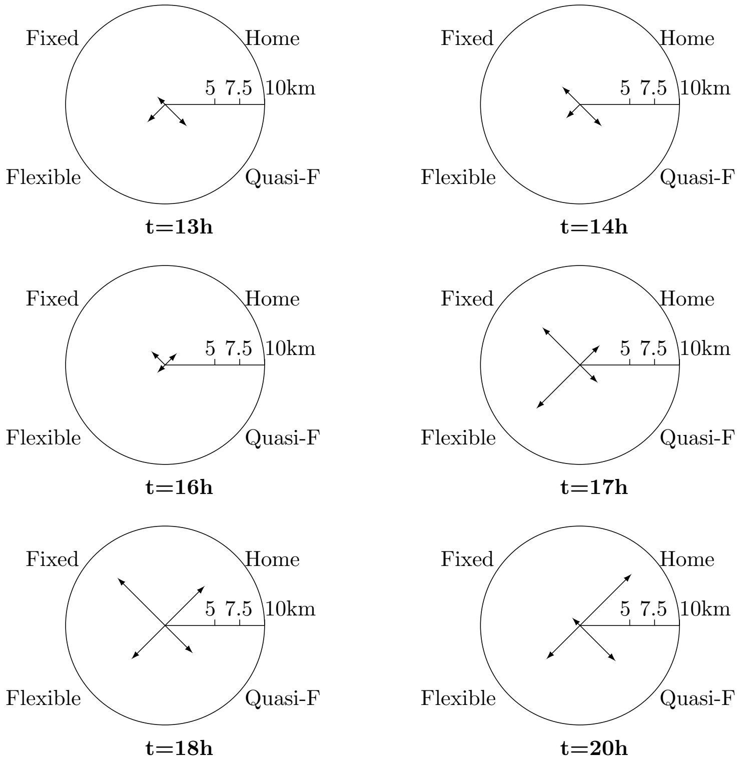


FIGURE 3.8.: 2D representation of the distance willing to be traveled by SM user #31 for participating at a specific activity type during lunch time and late afternoon retrieved from his Utility-Maximization model

After calculating the distances willing to be traveled for participating in different activity types during the day for all users, the CL divergence matrix is computed:

$$[CL] = \begin{pmatrix} CL_{0,0} & CL_{0,1} & \cdots & CL_{0,64} \\ \vdots & \vdots & \ddots & \vdots \\ CL_{j,0} & CL_{j,1} & \cdots & CL_{j,64} \\ \vdots & \vdots & \ddots & \vdots \\ CL_{31,0} & CL_{N-1,1} & \cdots & CL_{N-1,64} \end{pmatrix} \tag{3.14}$$

revealing users' willingness to travel similar distances for participating in certain types of activities as it is presented in Fig.3.9.

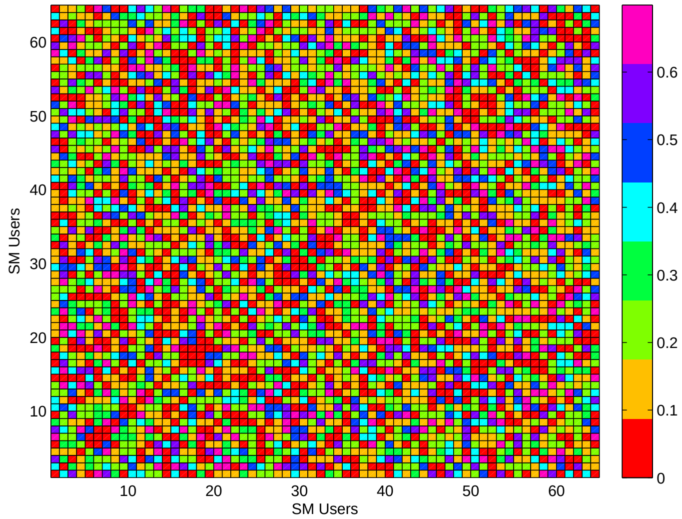
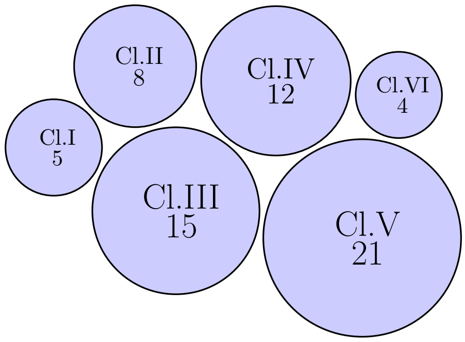


FIGURE 3.9.: Representation of the 65 × 65 symmetric CL Probability Distance for each couple of users in the form of a 2D Contour

Having computed the CL divergence matrix, the modified DBSCAN Alg.1 is applied to cluster users. The generation of at least five clusters, $m = 5$, is set as a threshold criterion and Alg.1 ran several times until the threshold criterion was satisfied leading to the generation of 5 or more clusters. For performing the above, a computer program was developed in Python 2.7. and the running time was 59min and 12sec using a 2.5GHz CPU and 4GB RAM. The results of the clustering are presented in Fig.3.10 including the number of SM users contained in each cluster.

In a future step, defining a PRM for predicting individual's mobility patterns from user-generated data and a utility-maximization model for capturing user's willingness to travel certain distances for participating in different types of activities offers the possibility of suggesting a common meeting place for multiple users who are willing to participate simultaneously in a similar activity type and are in close proximity. In Fig.3.11, a single-point estimate simulation is performed to construct step-by-step the daily mobility schedule of SM users # 17 and # 22 for a typical weekday based on their PRMs. At this figure, the focus is on the simulated timeframe from 15p.m. to 18p.m. since more flexible activities are expected to occur during that period. Users are placed in one location based on the simulation results and they can travel different distances that are represented with the diameter of a circle in order to participate in different types of activities according to their utility-maximization models. In Fig.3.11 the probability of performing a joint activity for users # 17 and # 22 is presented. Therefore, by simple inspection one can observe a feasible joint activity which has a flexible type and occurs at 17p.m.

The PRM model captures users' willingness to travel certain distances for participating in different types of activities for different day times and types (an example of implementing the model was presented in Fig. 3.7,3.8). The model learns automatically the users' habits from historic data and offers valuable insights that can be used as source of information for suggesting common activities to multiple users. For



Run	Clusters	m
I	3	5
II	2	5
III	4	5
IV	6	5

FIGURE 3.10.: Modified DBSCAN runnings until satisfying the threshold criterion and number of users at each generated Cluster for Running #IV

performing such action, users were clustered to capture the similarities on their mobility and activity patterns along with their willingness to travel similar distances to participate in certain types of activities (Fig.3.9). In Fig.3.9 was shown that most SM users have a CL distance below 0.5 while most users are concentrated in the range 0-0.2; however, there are several outliers.

In the end, users were clustered in 6 clusters after setting 5 clusters as a threshold containing from 4 to 21 users. The variance on clusters' size depends on the volume of concentrated users in the range 0-0.2. By simulating individuals' daily schedule with a step-by-step single-point estimate approach, one can identify automatically locations for performing joint activities as it is presented in Fig.3.11. This approach can be used for developing an application in order to suggest automatically the time and the location for performing a joint activity among users (i.e., restaurant which is located in the joint region).

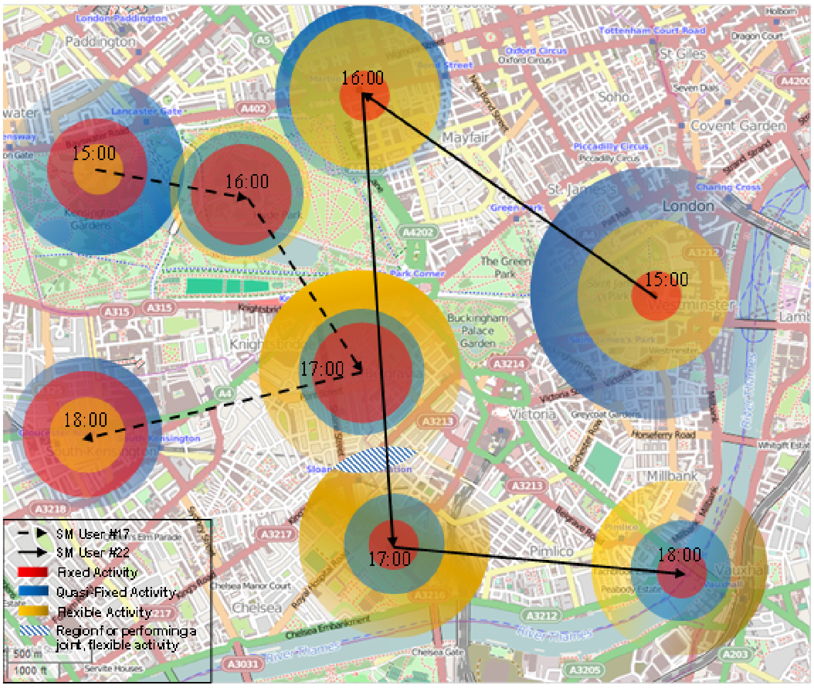


FIGURE 3.11.: Part of the mobility traces simulation with the use of single point estimates of users # 17 and # 22 from 15 to 18 p.m. Users' utility-maximization models show how far they are willing to travel to participate in different activity types given the location and time. At 17p.m., it is feasible to perform a joint, flexible activity

Optimizing Joint Leisure Activities' Locations & Starting Times

4.

The complexity of organizing joint leisure activities emerges from the need of satisfying simultaneously several activity participants with different preferences. This complexity is increased in urban environments with numerous places of interest due to the abundance of alternatives. For instance, if a group of x individuals with some form of social ties decides to organize a joint leisure activity, the selection of place and time will be based solely on the developed consciousness of individuals from prior experiences of attending leisure activities. Consequently, the solution space of alternative leisure activity places and starting activity times is not fully explored since individuals are not able to comprehend and re-call the whole set of alternatives for searching an optimal solution. With numerous groups of people organizing joint leisure activities on a daily basis, the problem scales up at a city level and the organization of joint leisure activities based solely on individuals' accumulated experiences leads to sub-optimal selection of activity destination and arrival times.

The importance of optimizing joint leisure activities is higher in urban environments since up to 60% of the conducted trips are related to leisure activities and the complexity of transport, social and activity networks leaves more room for optimization. For instance, as already noted, [1] posed that 29.2% of all daily trips are related to leisure activities, while 28% were conducted for shopping and personal business and 10.7% for other activities including escort. Similar results were observed on the New York Regional Travel survey [2] .

[2]: (2014), *New York Regional Travel Survey*

Selecting a location and individuals' arrival times for joint activity participation among multiple individuals is not a trivial task due to the problem of scalability and the lack of knowledge regarding users' preferences.

The real-time optimization of a joint leisure activity includes the selection of the location of the joint leisure activity, the starting time of the activity and the arrival time of all activity participants at that location. This requires the maximization of the perceived utility of all activity participants and depends on the a) current locations of individuals; b)the time of the day; c) the disutility of traveling from a current location to the location of the joint activity; d) the arrival time of each user to the joint activity location and the waiting time until the activity starts. The final suggestion of the location and starting time of a joint leisure activity can be provided at each individual via a smartphone application (refer to Fig.4.1).

The optimization of joint leisure activities requires strong analytics for identifying and modeling users' preferences from historical user-generated data and scalable optimization techniques for the suggestion

FIGURE 4.1.: Suggesting location and time of a joint leisure activity via mobile Apps

of the activity location and the arrival times in real-time. In the previous chapter, a utility maximization model derived from social media data (analysis of historical data from more than 4 months for each individual with automated pattern recognition techniques) for representing users' willingness to travel a certain distance for participating in leisure activities at different times of day was introduced. This needs though to be linked with a scalable optimization solution method for suggesting the preferred location of a joint leisure activity and the arrival times of individuals in near real time.

4.1. Users' willingness to travel in order to participate in leisure activities

New data feeds are handled automatically by the pattern recognition model described in the previous chapter yielding an improvement on capturing users' patterns.

That probabilistic model is adopted for describing the mobility patterns of an individual considering both the location and the activity dimensions. That probabilistic model was formed after a learning phase with the use of user-generated historical data and is user-specific:

$$P(k,t,L_m,A_n) = \frac{N(k,t,L_m,A_n)}{\sum_{A_n \in A} \sum_{L_m \in \Lambda} N(k,t,L_m,A_n)} \tag{4.1}$$

, where $N(k,t,L_m,A_n)$ is the number of times user k was in location L_m and participated in an activity A_n at time t.

Matrix P has $[k \times t \times L_m \times A_n]$ elements and represents the probability of a user, k, to be in a location and participate at an activity for different times and different day types. Those probabilities are calculated after accumulating the mobility footprint of an individual over time.

The transition probability over two consecutive time instances was also defined as:

$$a_{i,j} = P(S_{t+1} = q_j | S_t = q_i) \tag{4.2}$$

where the transition probability returns the probability to transfer from state q_i to state q_j at time t (refer to Fig.4.2). Eq.4.2 derives the probability of transition between two states $S_t = q_i \rightarrow S_{t+1} = q_j$ where each state is described by a location/activity/time combination. States

S_t, S_{t+1} represent the location and performed activity of an individual at a specific time instance and the same location/activity pair can be observed over two consecutive time instances of the same day. In that case, the individual evolves from one state to another but both states have the same activity/location pair attributes ($q_i = q_j$).

In the scenario of joint leisure activity participation, several participants should arrive at the joint activity location within a reasonable time window for avoiding excessive delays. For this reason, the state evolution timestep should be in many cases shorter than the approximate length of the typical activity in order to be synchronized with the possible arrival times of different travelers arriving at the joint leisure activity. In those cases, the shorter timestep can lead to a state evolution to a new state that has the same location/activity pair ($q_i = q_j$) as described above.

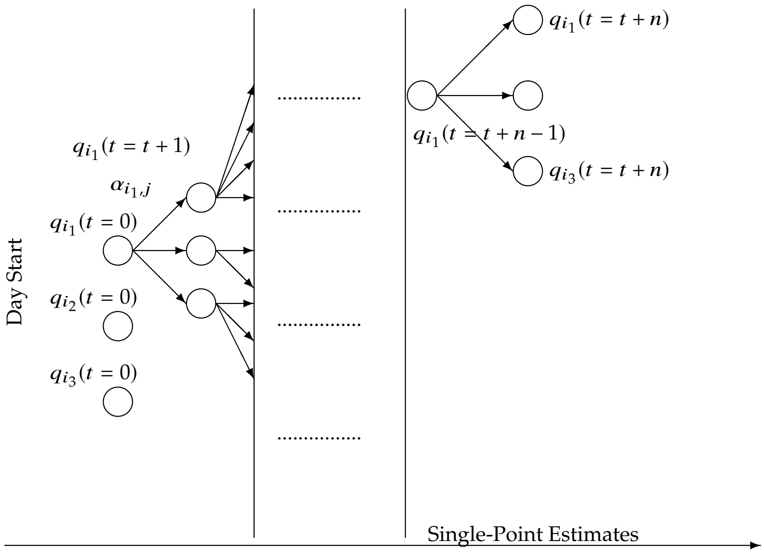


FIGURE 4.2.: Estimating the Daily Evolution of States over a day with the use of the probability matrix $P(k, t, L_m, A_n)$ and the transition probability $a_{i,j}$

In an attempt to model the decision-maker's choice for moving from his/her current state and travel a certain distance for participating in leisure activities, a utility function is defined. To decompose the activity selection model, the available choices are indexed by $j = \{1, ..., J\}$ where F_j is the traveled distance between the current location of the user and the choice j.

$$j = \begin{cases} 1 : F_j \leq x_0, & x_0 : \text{distance in km} \\ 2 : x_0 < F_j \leq x_1, & x_1 : \text{distance in km} \\ ... \\ J : F_j > x_{J-2}, & x_{J-2} : \text{distance in km} \end{cases}$$

Individual's utility $U_{tj}(k)$ from traveling a distance j for participating at a leisure activity ($A_t(k)$ = leisure activity) at time t is further defined as presented at the previous chapter. $D_{tj}(k, l)$ is also introduced as a new variable and denotes the dissatisfaction of user k when he has to wait for l minutes until all other participants arrive at the location of the joint activity. Disutility $D_{tj}(k, l)$ represents the degree of individual's dissatisfaction and is related to the decisions of other participants at the joint leisure activity regarding their arrival time.

The estimated utility function for participating at a leisure activity given the current location of on user, the suggested location of a joint leisure activity, the time of the day, and the disutility of individuals while waiting l minutes at the activity location for all other participants to arrive are utilized for solving the joint leisure activity optimization problem.

4.2. Joint Leisure Activity Optimization

Problem Description

Let assume that a set of users with social ties, N_E, would like to participate at a joint leisure activity. At the starting time t all individuals, $k \in N_E$, are at different locations. Let also assume that there is a set of candidate locations, N_S for joint leisure activities. Choosing an optimal location and time for the joint leisure activity requires the search of a broad spatio-temporal solution space where the perceived utility of individuals and the disutility due to the excess waiting time affect the solution.

For instance, if all individuals, $k \in N_E$, decide to participate at a joint leisure activity at location $\Lambda \in N_S$ and their arriving times to that location are $t_k \in N_f$, where N_f is a temporal set of possible arrival times, then the overall perceived utility and disutility for participating in such activity is:

$$f(U_{t_k\Lambda}(k), D_{t_k\Lambda}(k,l)) \tag{4.3}$$

, where $k \mid A_{t_k}(k)$ = leisure activity, $t_k \in N_f$ is the arrival time of each individual k at the location of the activity and $D_{t_k\Lambda}(k,l)$ the disulity of each individual k given that he/she has to wait l minutes before the activity starts. For each potential location of a joint leisure activity, Λ, the perceived participation utility and disutility at a particular time of day is represented in eq.4.3. Initially, in eq.4.3 all individuals, k, are considered. However, the probability of several individuals to perform a leisure activity at location Λ is equal to zero given their historical spatio-temporal patterns. In addition, other individuals might visit location Λ but not perform leisure activities there. Thus, the condition ($k \mid A_{t_k}(k)$ = leisure activity) is added in eq.4.3 for considering only the utility of individuals who perform a leisure activity $A_{t_k}(k)$ at time t_k in location Λ and not the utility of all others who are not visiting location Λ or performing other activity types there.

From the objective function, $\Theta = f(U_{t_k\Lambda}(k), D_{t_k\Lambda}(k,l))$, the optimal location and arriving times of each individual for participating at the joint leisure activity are determined (refer to Fig.4.3):

$$arg \max_{t_k \in N_f, \Lambda \in N_S} f(U_{t_k\Lambda}(k), D_{t_k\Lambda}(k,l)) \tag{4.4}$$

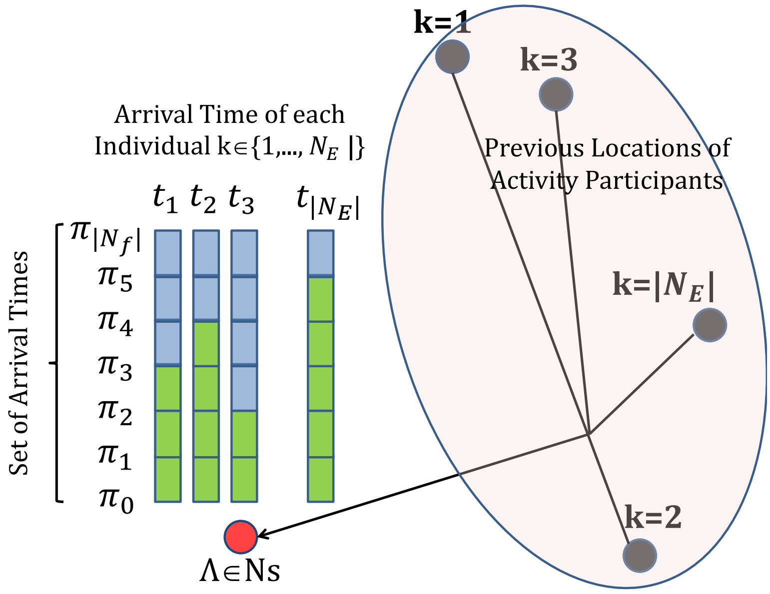


FIGURE 4.3.: Schematic View of the Search for the optimal location $\Lambda \in N_S$ and arrival times of all individuals $t_k, \forall k \in N_E$ for the next Joint Leisure Acivity

To solve this optimization problem, the computation of the effect of different arrival times $\in N_f$ to the disutility of users due to the excessive waiting times and the perceived utility of all individuals while attending a leisure activity candidate location requires:

$$| N_S | \times | N_f |^{|N_E|} \tag{4.5}$$

computations with base computational cost C, where $| N_S |$ is the cardinality of the set of location candidates for a leisure activity N_S, $| N_f |$ is the cardinality of the set of available arrival times of each user at the location of the activity and $| N_E |$ the number of the participants at the joint activity. The computational cost of calculating the global optimum solution grows exponentially with the number of individuals and the cardinality of the set, N_f, from which the arrival times can be selected; hence, the system fails to scale up -in its early phase already- as it is shown in Fig.4.4 where more than 1 hour is required for $| N_f |> 3$. This leads to the inability of optimizing the location and the arrival times of all participants at a joint leisure activity in practice even in small-scale scenarios.

Eq.4.4 returns the location for performing a joint leisure activity and the arrival times of each individual given their current locations and their perceived utility/disutility functions. The selection of the joint leisure activity location is not considering the future activity chains of the joint leisure activity participants. However, not considering the future activity plans of the joint leisure activity participants might lead to a significant deviation from their daily schedules; thus, reducing individuals' acceptance level of the suggested joint leisure activity.

To confront the above issue, the effect of selecting a joint leisure activity location based not only on the current locations of activity participants but also on their future state evolution probabilities is modeled in

Eq.4.6.

$$
\begin{aligned}
& arg \max_{t_k \in N_f, \Lambda \in N_S} f(U_{t_k\Lambda}(k), D_{t_k\Lambda}(k,l)) \\
& \text{subject to:} \qquad \sum_{q=1}^{|q|} P(k, t_k^*, q \mid k, t_k, \Lambda, A_n) \geq \vartheta \\
& \qquad\qquad q \neq (\Lambda, A_n)
\end{aligned}
\tag{4.6}
$$

where $\sum_{q=1}^{|q|} P(k, t_k^*, q \mid k, t_k, \Lambda, A_n)$ is the probability of each leisure activity participant k to move at any other location/activity pair $q \in \{1, ..., |q|\}$ and $q \neq (\Lambda, A_n)$ during his/her state evolution at time t_k^* given that he/she departed from the location/activity pair (Λ, A_n) of the joint leisure activity at time t_k.

If the probability of transition at a different location/activity pair for an activity participant k is higher than a threshold value $\vartheta \in (0, 1]$, then the selected location/activity pair (Λ, A_n) for the joint leisure activity does not constrict the activity participant from moving to another location/activity pair and continuing his/her daily activity plan. Therefore, the location/activity pair (Λ, A_n) is a feasible joint leisure activity option for participant k. This constraint guarantees that only location/activity pairs (Λ, A_n) that allow the joint leisure activity participants to continue their daily activity schedules are selected. Threshold ϑ refines the optimization solution space and if one sets a high threshold value ($\vartheta \simeq 1$), then he/she ensures that the selected joint leisure activity pair (Λ, A_n) will provide more alternative future activity options to the activity participants and the joint leisure activity suggestion will have more chances to be unanimously accepted.

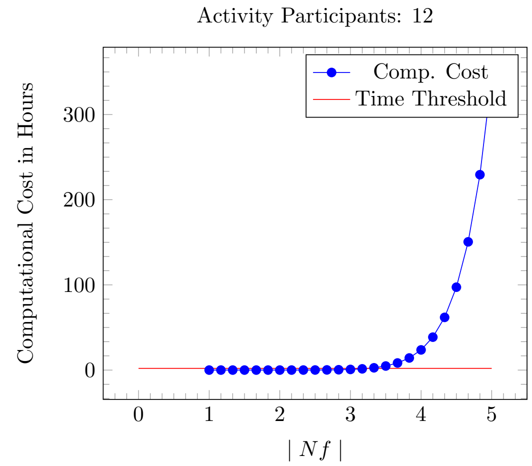


FIGURE 4.4.: Scalability Test in a Simplified Use Case with 12 activity participants and 60 leisure activity location candidates. The computational cost of calculating the objective function via a simplified model was $C = 0.084675 milisec.$ on a $2556 MHz$ processor machine with $1024 MB$ RAM. The number of alternative arrival time options that each individual can select from, $|N_f|$, is the horizontal axis variable

Stochastic Annealing Search

To address the scalability problem, a set of enhancements that reduce the complexity of the problem while ensuring that the final set of solutions is close to the optimal one are introduced. The main enhancement is the selective choice of arrival times from the set N_f for each individual for limiting the solution space via a "stochastic annealing" search method.

Following this approach, one reaches a final solution which is close to the optimum of the objective function Θ within a reasonable number of computations $\mid N_S \mid \times \mid N_f' \mid^{\mid N_E \mid}$ instead of $\mid N_S \mid \times \mid N_f \mid^{\mid N_E \mid}$; thus reducing the selection attempts on the set of arrival times $\mid N_f' \mid \lll \mid N_f \mid$.

At an initial search step, all candidate joint activity locations $\Lambda \in N_S$ are identified. Let assume that one candidate location $\Lambda \in N_S$ is examined. Starting from the first individual in the set N_E and following a successive order, Θ is optimized taking into account the arrival times of each individual. Then, the same approach is repeated at the next examined location until the exhaustion of all locations in N_S. Nevertheless, the computational cost is still not scalable if the set of arrival times that each individual can select from has more than 3 elements as shown in Fig.4.4.

For performing the problem-specific "stochastic annealing" search, the objective function, Θ, is computed with the assumption that all individuals arrive at the joint activity location at the fastest time possible given their current location ($t_k = t, \forall k \in N_E$). In the initial search, the first individual is allowed to arrive at the location of the leisure activity at time $t_k \in N_f$ which optimizes the Θ. For performing this action, $\mid N_f \mid$ computations are required. If one repeats the same for the second individual assuming that all other individuals arrive to the activity location at the earliest possible time, the cost is again $\mid N_f \mid$ and if this is repeated for all individuals at all locations the number of computations is $\mid N_f \mid \times \mid N_E \mid \times \mid N_S \mid$.

Let assume that the available set of arrival times, N_f, of one individual k at a leisure activity location is $N_f = \{\pi_0, \pi_1, ..., \pi_n, ..., \pi_x, ..., \pi_{|N_f|}\}$ where π_0 is the earliest possible arrival with a computed performance cost Θ_{π_0}. π_x is also the arrival time from the initial search that optimized the performance cost Θ_{π_x} given that all other individuals arrive at the location as fast as possible. The stochastic annealing search starts by picking the next arrival time from the N_f set of values for ensuring that the solution space is efficiently explored given that the number of available selections is only $\mid N_{f'} \mid \leq 3$ where $\mid N_{f'} \mid \lll \mid N_f \mid$.

At a first selection step, the arrival time $\pi_{\kappa*} \in N_f$ with the highest selection probability based on the already computed performance of π_0 and π_x is picked. $\pi_{\kappa*}$ has the highest probability of selection among all other arrival times $P(\pi_{\kappa*}) \geq P(\pi_\kappa)$, $\pi_\kappa \in N_f$. Then, the $\Theta_{\pi_{\kappa*}}$ is computed (refer to Fig.4.5). The selection probability of any control measure $\pi_\kappa \in N_f$ is defined as:

$$P(\pi_\kappa) = \frac{(\Theta_{\pi_0}/\Theta_{\pi_x})\pi_\kappa + L}{\sum_{\pi_\kappa \in N_f}(\Theta_{\pi_0}/\Theta_{\pi_x})\pi_\kappa + L} \tag{4.7}$$

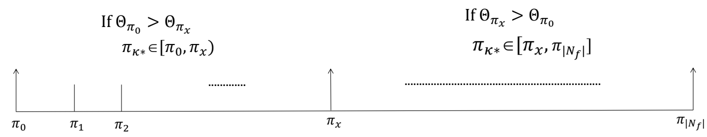


FIGURE 4.5.: Stochastic Annealing Selection of Arrival Time from N_f

where

$$L = W||(\pi_\kappa - \pi_0)(\pi_\kappa - \pi_x)/|\ N_f\ |\ || \tag{4.8}$$

and the term $W||(\pi_\kappa - \pi_0)(\pi_\kappa - \pi_x)/|\ N_f\ |\ ||$ has a higher value when the arrival time π_κ is closer to other arrival times for which the Θ is already calculated (i.e., π_0 and π_x). W is a weight factor that increases the probability of selecting an arrival time which is closer to the ones that there is information regarding their performance.

At a second selection step, the probabilities of selecting the next arrival time $\pi_{\kappa**} \in N_f$ are redefined based on the already computed Θ_{π_0}, Θ_{π_x} and $\Theta_{\pi_{\kappa*}}$ Let assume that $\pi_0 \leq \pi_{\kappa*} \leq \pi_x$. Then, for all $\pi_\kappa \in N_f$ if $\pi_\kappa \leq \pi_{\kappa*}$:

$$P(\pi_\kappa) = \frac{(\Theta_{\pi_0}/\Theta_{\pi_{\kappa*}})\pi_\kappa + W||(\pi_\kappa - \pi_0)(\pi_\kappa - \pi_{\kappa*})/N_f||}{\sum_{\pi_\kappa \in N_f}(\Theta_{\pi_0}/\Theta'_{\pi_x})\pi_\kappa + L} \tag{4.9}$$

Else if $\pi_\kappa \geqslant \pi_{\kappa*}$:

$$P(\pi_\kappa) = \frac{(\Theta_{\pi_{\kappa*}}/\Theta_{\pi_{x'}})\pi_\kappa + W||(\pi_\kappa - \pi_{\kappa*})(\pi_\kappa - \pi_x)/N_f||}{\sum_{\pi_\kappa \in N_f}(\Theta_{\pi_n}/\Theta'_{\pi_x})\pi_\kappa + L} \tag{4.10}$$

After that, $\pi_{\kappa**}$ such that $P(\pi_{\kappa**}) \geq P(\pi_\kappa), \forall \pi_\kappa \in N_f$ is chosen and the $\Theta_{\pi_{\kappa**}}$ is computed.

Finally, one more step follows and the outcome of the constrained solution search returns a first direction towards the Θ optimum. During the computation, one arrival time $\in N_{f'}$ is selected for each individual at each examined activity location as it is summarized in Alg.2.

As shown more clearly in Alg2, the set of arrival time selection candidates $\{\pi_{\kappa*}, \pi_{\kappa**}, \pi_{\kappa***}\}$ is defined with the stochastic feature of the annealing search before start applying simulated annealing at the reduced search space $N_{f'}$ for finding an approximation to the global optimum solution. At the stochastic step, the set of $N_{f'} = \{\pi_{\kappa*}, \pi_{\kappa**}, \pi_{\kappa***}\}$ solution candidates for each activity participant which shows his/her strongest preferences in terms of activity arrival times is pre-selected out of the entire N_f set and the simulated annealing search is focused only at search space $N_{f'}$; thus, truncating the size of the exploration space.

This is the key differentiator of the proposed stochastic annealing method; since, simulated annealing is applied at a search space with high chances to contain a solution close to the global optimum. Otherwise, if simulated annealing was applied in the entire set N_f without considering the stochastic set-selection feature, it would not be able to explore efficiently the vast solution space for finding an appropriate approximation to the global optimum given the limitation of using only 3 search iterations due to the scalability issue described in Fig.4.4.

Algorithm 2: Stochastic Annealing Search with $| N_S | \times | N_{f'} |^{|N_E|} + | N_S | \times | N_f | \times | N_E |$ computations

for *all candidate locations* $\Lambda \in N_S$ **do**
  Compute Θ when all individuals arrive at Λ as fast as possible;
  for *each individual* $k = \{1, ..., | N_E |\}$ **do**
    Select the optimal arrival time $t_k \in N_f$ for individual k:Θ_{π_x} considering that all other individuals arrive at Λ as fast as possible;
  end
  for *each individual* $k = \{1, ..., | N_E |\}$ **do**
    for *each arrival time* $\pi_\kappa = \{\pi_0, ..., \pi_{|N_f|}\}$ **do**
      Calculate $P(\pi_\kappa(k))$;
      Select $\pi_{\kappa*}(k)$ such that $P(\pi_{\kappa*}(k)) \geq P(\pi_\kappa(k))$;
      Recalculate $P(\pi_\kappa(k))$ and select $\pi_{\kappa**}(k)$ such that $P(\pi_{\kappa**}(k)) \geq P(\pi_\kappa(k))$;
      Recalculate $P(\pi_\kappa(k))$ and select $\pi_{\kappa***}(k)$ such that $P(\pi_{\kappa***}(k)) \geq P(\pi_\kappa(k))$;
    end
  end
  for $\pi_{\kappa*}(1)$,$\pi_{\kappa**}(1)$,$\pi_{\kappa***}(1)$ **do**
    for $\pi_{\kappa*}(2)$,$\pi_{\kappa**}(2)$,$\pi_{\kappa***}(2)$ **do**
      .....;
      for $\pi_{\kappa*}(N_f)$,$\pi_{\kappa**}(2)$,$\pi_{\kappa***}(| N_E |)$ **do**
        Calculate $New.\Theta$;
        if $New.\Theta > \Theta$ **then**
          $\Theta = New.\Theta$;
        end
      end
      .....;
    end
  end
end
Return the selected arrival time for each individual and the location of the joint leisure activity;

4.3. Computing the Location and Starting time of Joint Leisure Activities

In practice, the user-generated data from 75 Smatrphones that post on Twitter and the Utility-maximization model for each user was generated representing the distance he/she is willing to travel for participating in different activity types during the day. A short summary of the generated Utility-maxization models is presented in Fig.4.6 where only the maximum distance that one user is willing to travel

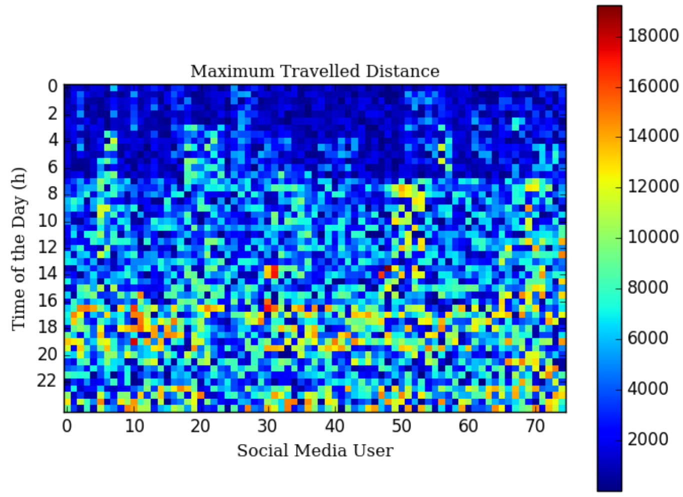


FIGURE 4.6.: Plot of 75 Users' Maximum Travelled Distance (in meters) for participating at leisure activities at different times of day. The utilized user-generated data comes from 75 Smatrphones that post on Twitter

for participating at a leisure activity at different times of day is plotted due to the immense amount of data.

The stochastic annealing optimization algorithm is tested after retrieving the Utility-based models of each individual from analyzing their social media data. The objective of the joint leisure travel optimization technique is to propose an activity location and the arrival times to each individual participating at the activity given their current location. As described previously, the brute-force optimization requires an exponential number of computations $| N_S | \times | N_f |^{|N_E|}$ of the objective function $\Theta = f(U_{t_k \Lambda}(k), D_{t_k \Lambda}(l,k))$ for retrieving the optimal value.

At this point, it should be mentioned that Θ is not a fixed function. For instance, the value of Θ for an examined activity location and set of arrival times can be estimated via a micro or macro-level simulation which offers higher granularity or via a simpler equation which associates Θ with the parameters of the examined location and the arrival times. For the implementation of the stochastic annealing algorithm, the following function for calculating the objective function is used:

$$\Theta_{\Lambda,t_1,t_2,\ldots,t_{|N_E|}} = \frac{\sum_{k=1}^{|N_E|} U_{t_i,\Lambda}(k)}{| N_E |} - w_k \frac{\sum_{k=1}^{|N_E|} D_{t_k,\Lambda}(k, T_p - t_k)}{| N_E |} \tag{4.11}$$

,where $t_k \in N_f$, T_p the starting time of the joint leisure activity and w_k a weight factor for individual k. Eq.4.11 includes the utility and disutility of all individuals k at location Λ. However, only the utility/disutility of individuals who perform a leisure activity $A_{t_k}(k)$ at time t_k in location Λ should be considered in the objective function. Therefore, the condition ($A_{t_k}(k)$ = leisure activity) is added as an additional constraint of eq.4.11.

The proposed function serves as a reference and different functions or more detailed simulation-based approaches can be utilized. Nevertheless, in any case, a computational cost of C is required for estimating

Θ for a given location and a set of arrival times; thus, requiring a computational cost of $C \times \mid N_S \mid \times \mid N_f \mid^{|N_E|}$.

Departing from this point, optimizing the leisure activity locations and the arrival times of 75 social media users solves an unrealistic problem since such activity participation levels of users with social ties occur only in the rare case of special events (i.e., sport events, pre-organized social events) where the location of the activity is already known. To tackle this issue, one needs to use the generated clusters from the examined users from the previous chapter based on the observed distance between the utility-mazimization models of users over time. This distance represented the similarity of users' willingness to travel similar distances for participating in leisure activities.

For all users in each cluster, their daily mobility/activity plans based on their evolution of states described in Fig.4.2 were derived from Eq.4.1,4.2 at the learning phase. The states of all users are evolving in a rolling horizon and if at some point in time all users that belong to the same cluster are at the same location and this is a leisure activity type location, it is assumed that this is a Joint Leisure Activity Instance. Then, the values of the activity starting time, the activity location and the arrival time of each individual are stored and the perceived utility of the users that participated to the joint leisure activity is calculated.

While the evolution of states continuous, the users of one cluster can meet again at another location which is again the location of a leisure activity; thus, assuming that this is a 2nd joint leisure activity instance. The evolution of states for each individual can continue until exhaustion; however, at an initial stage of analysis, the state evolution procedure is terminated for all users that belong in a cluster when the occurrence of nine (9) joint leisure activity Instances is observed.

The nine observed Instances of Joint Leisure Activities at each cluster have occurred without external interference and form the basic scenario ("do-nothing" scenario). The values of the perceived utility of all clusters for the "do-nothing" scenario at each Instance of Joint Leisure Activity occurrence are stored in Table 4.1.

Optimizing each joint leisure activity for each cluster at each Instance of activity occurrence requires $\mid Ins. \mid \times \mid Cl \mid \times \mid N_S \mid \times \mid N_f \mid^{|N_E|}$ computations where $\mid Cl \mid = 10$ is the number of generated clusters, $\mid Ins. \mid = 9$ the total Instances of joint activity occurrence and N_S the number of alternative joint activity locations. In addition, N_f is the set from which one can select different arrival times to the location of the joint activity and it is formed based on the assumption that the participant who arrives first cannot wait more than 80 minutes until the arrival of the last participant due to the inconvenience caused. For this test case, N_f contains 10 values; therefore for each user k, the arrival time $t_k \in \{1, ..., \mid N_f \mid = 10\}$.

The proposed Stochastic Annealing heuristic was tested against the exhaustive enumeration method that required $\mid Ins. \mid \times \mid Cl \mid \times \mid N_S \mid \times \mid N_f \mid^{|N_E|}$ computations in the case of $\mid N_f \mid = 10$, $\mid Ins. \mid = 9$ and $\mid Cl \mid = 10$ clusters containing up to 9 users on a 2556MHz processor machine with 1024MB RAM and the computational costs of

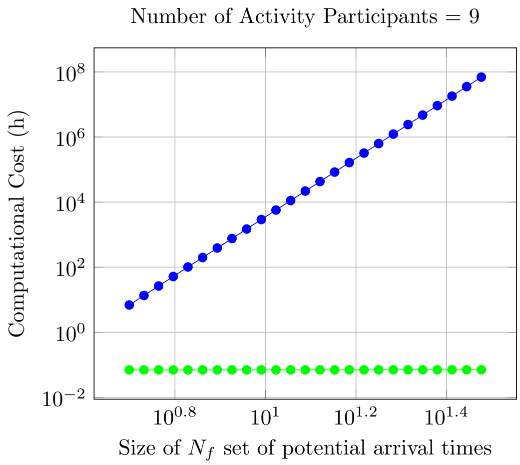


FIGURE 4.7.: Comparison of Computational Cost between Brute Force and Stochastic Annealing Search. The Plot is in Logarithmic Scale

both methods are shown in Fig.4.7. Fig.4.7 shows that the proposed algorithm computes the optimization solution in linear scale compared to the exponential behavior exhibited by the brute force approach and, most importantly, the computational cost stays at lower levels that allow the implementation of real-time applications.

The stochastic annealing method finds approximations of the global perceived utility optimum for each cluster of users at each joint activity instance. The global perceived utility optimum is defined with respect to the group of individuals who have the potential to participate at each joint leisure activity instance (decentralized approach) and not with respect to the network conditions. After running the algorithm for all test cases, new suggestions regarding the location of a joint activity and the arrival times of all users at each cluster are calculated. The optimization findings are suggested to users and their perceived utility from participating at joint leisure activities at each instance is calculated and presented in Table 4.1 and Fig 4.8. Table 4.1 and Fig 4.8, show significant increase on the perceived utility of users that varies from up to 1.5 to up to 5 times compared to the observed utility satisfaction from the do-nothing scenario.

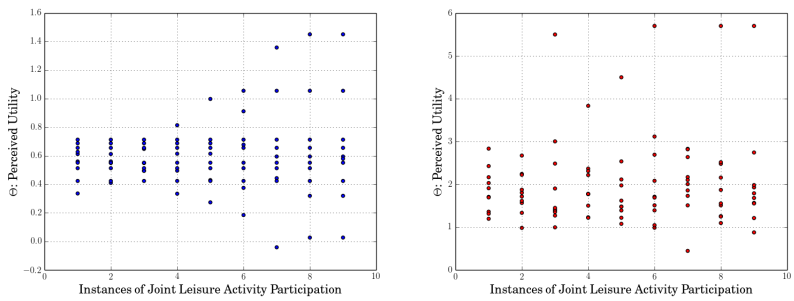


FIGURE 4.8.: Perceived Utility values for all clusters before [left] and after [right] the Stochastic Annealing Optimization

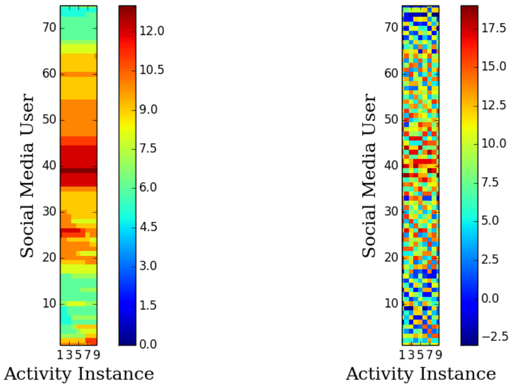


FIGURE 4.9.: Arrival Times of each Individual at each joint activity location for each one of the nine instances. The color bar shows the difference between the arrival time of one user and the arrival time of the next user who arrived at the same location

Table 4.1.: Perceived Utility of each Cluster of Joint Activity Participants at different times before, Θ_{real}, after the optimization of the activity Location and the Arrival Times, $\Theta_{opt.}$, and with brute-force Θ_{brute}

Cluster		Ins.$_1$	Ins.$_2$	Ins.$_3$	Ins.$_4$	Ins.$_5$	Ins.$_6$	Ins.$_7$	Ins.$_8$	Ins.$_9$
	Θ_{real}	0.61	0.61	0.65	0.81	1.00	1.06	1.36	1.45	1.45
1	$\Theta_{opt.}$	1.71	0.98	2.48	2.37	2.11	2.69	2.83	2.48	2.74
	Θ_{brute}	1.89	1.23	2.74	2.69	2.81	2.92	3.11	2.98	3.17
	Θ_{real}	0.34	0.41	0.55	0.61	0.61	0.68	0.59	0.59	0.59
2	$\Theta_{opt.}$	1.91	2.67	3.00	2.22	1.39	3.12	2.82	1.87	1.68
	Θ_{brute}	2.04	2.72	3.41	2.55	1.78	3.26	3.07	2.11	2.01
	Θ_{real}	0.56	0.56	0.50	0.50	0.43	0.38	0.44	0.32	0.32
3	$\Theta_{opt.}$	2.83	1.33	1.38	1.77	1.08	1.71	2.17	2.16	1.79
	Θ_{brute}	2.99	1.56	1.54	2.01	1.37	1.79	2.45	2.37	2.01
	Θ_{real}	0.63	0.55	0.49	0.33	0.27	0.18	-0.04	0.03	0.03
4	$\Theta_{opt.}$	2.17	1.80	0.99	1.23	1.22	0.99	0.44	1.51	0.88
	Θ_{brute}	2.34	2.02	1.17	1.56	1.42	1.59	0.46	1.62	0.96
	Θ_{real}	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71	0.71
5	$\Theta_{opt.}$	1.69	2.24	1.35	1.78	1.48	1.51	1.73	1.09	1.93
	Θ_{brute}	1.78	2.27	1.51	1.81	1.62	1.84	1.77	1.41	2.27
	Θ_{real}	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66	0.66
6	$\Theta_{opt.}$	1.36	1.57	1.41	1.77	1.97	1.69	2.01	1.55	1.21
	Θ_{brute}	1.51	1.93	1.46	2.01	2.07	1.87	2.03	1.67	1.42
	Θ_{real}	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55	0.55
7	$\Theta_{opt.}$	1.31	1.87	1.45	1.22	1.48	1.04	1.51	1.26	1.56
	Θ_{brute}	1.71	1.94	1.65	1.42	1.58	1.27	1.59	1.41	1.63
	Θ_{real}	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.51	0.58
8	$\Theta_{opt.}$	1.19	1.72	1.90	1.50	1.62	1.39	2.10	1.25	1.55
	Θ_{brute}	1.32	1.79	1.94	1.76	1.87	1.61	2.17	1.39	1.71
	Θ_{real}	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42	0.42
9	$\Theta_{opt.}$	2.43	2.22	1.27	2.31	2.54	2.08	1.86	2.52	1.98
	Θ_{brute}	2.45	2.29	1.42	2.39	2.63	2.17	1.99	2.56	2.07
	Θ_{real}	0.69	0.69	0.69	0.69	0.69	0.91	1.06	1.06	1.06
10	$\Theta_{opt.}$	2.03	1.61	5.50	3.83	4.50	5.70	2.64	5.70	5.70
	Θ_{brute}	2.11	2.27	5.50	4.47	4.96	5.70	2.91	5.70	5.70

For a more detailed insight on the performance of the stochastic annealing optimizer, Fig.4.9 shows the deviation of the arrival times of each couple of users at the location of the joint activity. A *couple of users* is the set of two users who belong to the same cluster and one arrives to the location of the joint leisure activity just before the other at the "do-nothing" scenario (users with successive arrival times). After the implementation of the optimizer, the deviation of arrival times changed abruptly and were also observed changes on the order of successive users when one user that arrived prior to another one at the "do-nothing" scenario now arrives after him/her. This is shown in Fig.4.9 where some user couples have negative values of arrival times deviation. The higher deviation of the arrival times of successive users $\{-3, ..., +18min.\}$ instead of $\{0, ..., +12min.\}$ at the "do-nothing" scenario case demonstrates the availability of valuable alternative options when searching a broad spectrum of the solution space in limited time and justifies the up to 5 times increase of the perceived utility (refer to the perceived utility of users from cluster 10 at Fig.4.10).

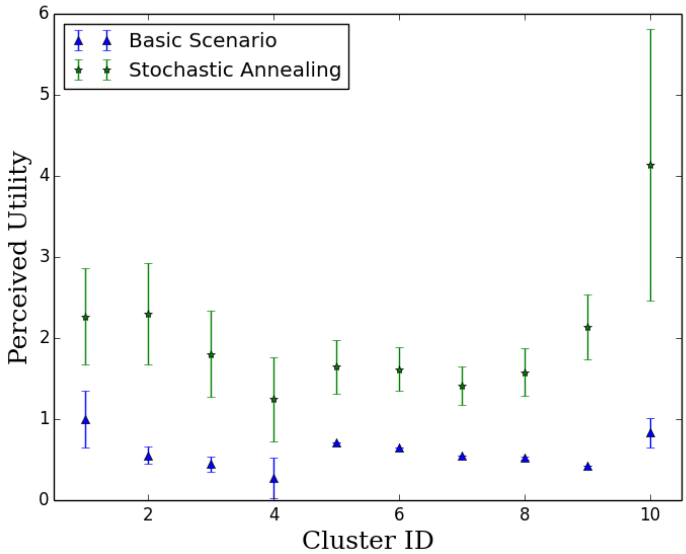


FIGURE 4.10.: Average Value and Standard Deviation of the Perceived Utility of each Cluster of activity participants over the 9-instance test case

Fig.4.10 summarizes the results and shows the average perceived utility of each cluster of users by averaging the perceived utility at each one of the nine Instances of joint leisure activity occurrence. After the optimization, cluster 10 had the most significant performance increase (up to 5 times) but also had the highest perceived utility deviation. This can be explained by the small size of the cluster (contains only 3 users). From all other clusters, there is a perceived utility increase of up to 2.5 times that demonstrates the potential benefits of optimizing leisure activities. At all cases, the optimized perceived utility had higher deviation than the do-nothing one. The higher deviation is partly explained from the selection of arrival times from an increased set of values that leverages the benefits of arrival times rescheduling; hence, in some cases extreme arrival time values are adopted while in other more conservative options were selected.

Finally, the approximation of the perceived utility optimum derived from the stochastic annealing optimization from the 1st and 2nd Joint Leisure Activity Instance is tested against other problem-customized optimization algorithms. The objective is to understand if the stochastic annealing optimizer converges fast enough to find a solution close to the optimum one; therefore, the main focus was on the 1st and 2nd Joint Activity Instance by running the stochastic annealing algorithm 12 times to understand the level of dispersion of the solution approximations that are computed at each run. Additionally, other problem-customized optimization algorithms were run 12 times at the same scenario for performing the comparative analysis.

Here it should be mentioned that there are no works in literature on optimizing leisure joint activities in a rolling horizon framework for real-time applications since the most related works belong to the area of ride-sharing optimization (refer to [73],[74]). Therefore, due to the absence of alternative heuristics, well-known algorithms that are used in the area of fleet management and belong to the algorithmic families of genetic algorithms and hill climbing were customized for approximating global optimums. To reduce the comparison bias, the customized hill climbing and genetic algorithm (GA) were also allowed to select $|Nf'|$=3 out of $|Nf|$=10 elements (i.e., 3 random start ascent selections in the case of hill climbing and 3 population evolution phases in the case of GA). Due to that, all three algorithms require similar number of computations.

The KPI of the performance of each algorithm is the improvement of the perceived utility of all users within clusters that participated at a joint leisure activity. Fig.4.11 shows the performance of each optimization algorithm after re-running it 12 times for approximating an optimal solution to the scenarios of the 1st and 2nd Instance of Joint Leisure Activity occurrence. The average value of the Perceived Utility of each cluster after the use of Stochastic Annealing, the Genetic Algorithm and the Hill Climbing method after aggregating the solution approximation results from the 12 re-runs are finally summarized in Fig.4.12.

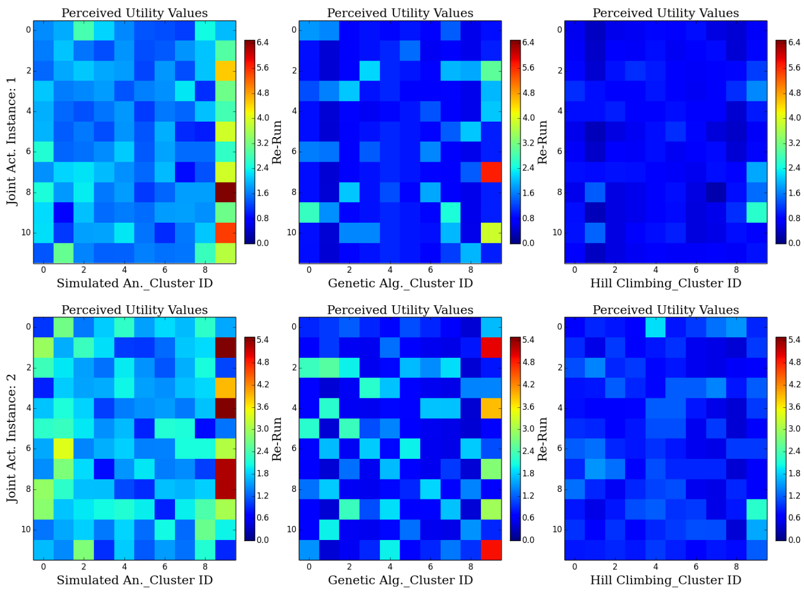


FIGURE 4.11.: Values of the Optimized Perceived Utility for each cluster after each re-run of the optimization algorithms (12 re-runs in total). Each re-run of the same algorithm finds another approximation to the global optimum and the vertical axis demonstrates the dispersion of the approximated solutions.

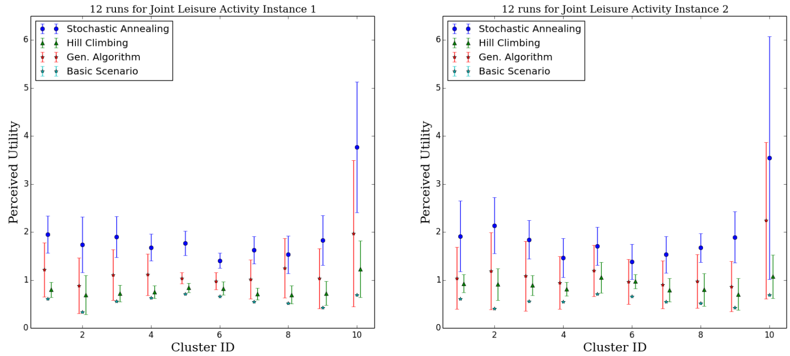


FIGURE 4.12.: Average Value and Standard Deviation of the Perceived Utility of each Cluster after the 12 re-runs of the Stochastic Annealing, the Genetic Algorithm and the Hill Climbing optimization algorithms for the scenarios of the 1st and 2nd Joint Leisure Activity Instance.

Due to the improved scalability, this approach can be used to develop Web and Smartphone applications for suggesting automatically the time and the location for performing a joint leisure activity. After applying the stochastic annealing search for optimizing 9 Joint Leisure Activity Instances, the perceived utility of users was increased up to 3 times compared to the basic scenario because of the rescheduling of the arrival times of individuals and the activity locations as presented in Fig.4.10,4.9.

From Fig.4.11,4.12, stochastic annealing found approximations to the optimal solution that increased up to 2 times the perceived utility of users in a cluster compared to the hill climbing and the GA. The Hill Climbing method had the most weak performance on approximating the global optimum while departing from the do-nothing scenario. It showed significantly low dispersion among the approximated solution values from the 12 re-runs and this is related to the inability to ascend to the top of the hill when there are only $| N_{f'} |= 3$ available iterations. In addition, many times the solution search started from the wrong hill and the algorithm was trapped in a local optimum without being able to jump to another hill for continuing its exploration.

Finally, the customized Genetic Algorithm oftentimes failed to identify an improved new generation and after the crossover phase many offsprings yield similar or worse solutions compared to their parents. However, GA showed a consistently better performance compared to the hill climbing method. In fact, due to the aggressive solution space exploration, at some cases its performance was close to the performance of the proposed stochastic search. The main problem though was the huge performance dispersion among re-runs at the same scenario. Given the threshold of 3 available population generations, there were several runs that the GA did not provide any improvements to the do-nothing scenario. Then, suddenly at one re-run the GA might manage to find a significantly improved solution due to the aggressive solution space search, even if such probability is low. Nevertheless, at most cases the approximated solutions offered limited added value to the activity participants and this was the key difference between the GA and the stochastic annealing method.

To conclude, the stochastic annealing method proposed consistently solutions close to the optimum without facing significant dispersion among re-runs, due mainly to the stochastic feature of pre-selecting a solution space with high chances of finding a close-to-optimum solution before start applying the simulated annealing search.

4.4. Potential of Problem Approximation with Centroids

The initial computational cost of solving the location and time optimization problem for each joint leisure activity was exponential and required a set of $\mid N_S \mid \times \mid N_f \mid^{\mid N_E \mid}$ computations. This set of computations required the implementation of heuristics given the large size of alternative joint leisure activity location options N_S in a city environment and the set of potential arrival times N_f.

However, a two-stage approximation method of selecting the activity location can be utilized for reducing the number of alternative activity locations N_f; thus, reducing the initial complexity of the optimization problem. In such approach, the city can be split in different areas with the use of centroids (let say $\mid N_C \mid$ centroids) where each centroid contains a number of locations $\mid N_{C_L} \mid$. In such approach, the first optimization stage can select the optimal centroid and activity time for a joint leisure activity with a required number of computations $\mid N_C \mid \times \mid N_f \mid^{\mid N_E \mid}$. After that, the search of the optimal activity location can be implemented within the centroid at a second-stage optimization.

This computational cost is significantly reduced due to the replacement of alternative locations, N_S, with alternative centroids, N_C. However, it still remains exponential and assuming the split of a city into 10 centroids the brute-force computational cost of the problem as it was defined in Fig.4.4 becomes the cost of Fig.4.13.

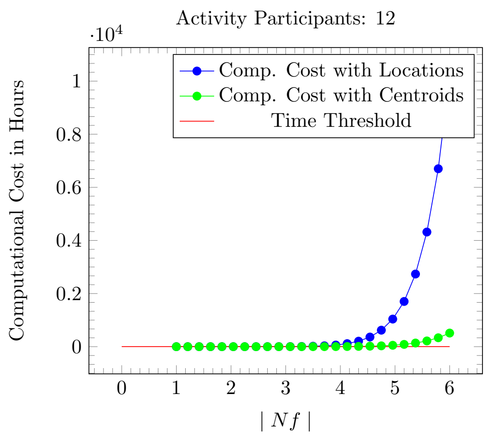


FIGURE 4.13.: Scalability Test in a Simplified Use Case with 12 activity participants and 200 leisure activity location candidates [blue dots] replaced by 10 centroids [green dots]. The computational cost of calculating the objective function via a simplified model was $C = 0.084675 milisec.$ on a $2556 MHz$ processor machine with $1024 MB$RAM.

From the above figure it is evident that the optimization cost is reduced significantly for more than 5 potential starting time options with the use of centroids. However, due to the exponential computational complexity, even with the use of centroids the problem requires the implementation of heuristics instead of simple enumeration for more than 4 potential starting time options. The conclusion of the above is that the problem complexity can be significantly reduced with the replacement of potential joint leisure activity locations with centroids, but that reduction is not enough for avoiding the utilization of solution approximations with heuristics.

Lastly, one should be cautious regarding the approximation with centroids because at the first optimization stage the optimal centroid is selected and that bounds the search of the optimal location within the limits of that centroid. The problem with the approximation with centroids is that a centroid is not a specific location but a large geographical area that entails a series of locations. For resolving this issue, the center of each centroid is selected as the representative of the entire centroid. Therefore, the distance between the current location of each activity participant and the center of the centroid is the key factor for selecting the optimal centroid. However, within the centroid itself the locations of the potential joint leisure activities might be far away from the center of the centroid which leads to an inaccurate approximation of the problem. To resolve this issue and provide a better approximation of the joint activity location optimization problem with the use of centroids, each centroid should be constructed in a specific way. Ideally, each centroid should: (a) contain a series of potential joint leisure activity locations which are in close proximity with each other yielding a low dispersion; (b) not contain several groups of potential joint activity locations which are far away from each other. In case that happens, those groups should be split and a new centroid should be generated for representing each group; (c) the center of the centroid should not be the geographical center, but the center of the potential joint leisure activity locations. By developing centroids following the above rules the inherent accuracy error of selecting the optimal joint leisure activity location with the use of centroid approximation is reduced, but there is no guarantee that there can finally be no accuracy loss.

5. Demand-Responsive Public Transport Re-scheduling

Mode selection and the increase of public transport mode ridership is the next challenge after the optimization of joint leisure activities' location and starting time. The work of [75] focused on the selection of transport modes for participating at activities from all members of a household by utilizing the concept of Random Utility Maximization (RUM) and adding constraints to the selection problem (i.e., private mode availability subject to the use from another member of the household). Although, the literature on dis-aggregate (e.g., [76]) or tour-based (e.g., [77]) mode choice is vast, the work of [75] is one of the closest prior arts since it tackles the problem of mode selection for a joint activity from several household members. However, the transport mode selection for a joint leisure activity has a higher complexity level than the household-based activity planning because: i) joint leisure activity participants start their trip from different origins and not from a common location (i.e., not necessarily home-based), ii) the arrival time punctuality at the location of the leisure activity matters since the arrival time variance of all activity participants should be low to avoid excess waiting before the activity starts.

The main challenge that arises is how to model and improve the joint-leisure-activity ridership of demand-responsive transportation via rescheduling the timetables of public transport modes subject to a set of operational constraints. For this, the problem of public transport service re-scheduling considering operational regulations and the quality of service is modeled and a scalable heuristic search method for dynamically re-scheduling the public transport schedules of demand-responsive systems and increasing their ridership related to joint leisure trips is introduced.

5.1. Public transport rescheduling for increasing the passenger ridership related to joint leisure activities

To increase public transportation ridership for trips related to joint leisure activities, the planned schedules of public transport modes should be adapted to the joint-leisure-activity travelers' requirements. Demand responsive public transportation (DRPT) has been viewed as a mean of passenger communication with the transport operator in a direct way allowing the transport operator to create custom routes based on a priori knowledge of the passengers' locations, destinations and schedules ([78] , [79]).

[78]: Mauri et al. (2009), 'Customers' satisfaction in a dial-a-ride problem'

For tackling the public transport rescheduling problem, a first assumption is that DRPT routes are fixed and only their timeschedules (i.e., departure times) can change subject to spatio-temporal demand variations related to joint leisure activities. The key individual-based data for deriving Joint Leisure activities demand variation in space and time is obtained, as discussed in previous chapters, from Smartphone apps that can open one-way communication channels between the travelers and the public transport operator (i.e., social media apps). In the past, smartphone apps have utilized for improving dynamic travel decisions [80] . In this thesis, Smartphone data from Social Media Apps is transmitted to the control center of the transport operator (either via direct transmission or via a data crawling campaign) and provides information of joint-leisure-activity demand for adapting the DRPT operations to it via changing the planned timeschedules.

[80]: Dickinson et al. (2014), 'Tourism and the smartphone app: Capabilities, emerging practice and scope in the travel domain'

Let assume that over the time period of a day a number of public transport trips $\pi = \{\pi_1, \pi_2, ..., \pi_\pi\}$ can serve joint leisure activity passengers. Those public transport trips belong, in general, to different public transport services while, some of them, can belong to the same public transport service (i.e., the bus service line). Let for instance a trip $\pi_i \in \pi$ to be part of a public transport service p. This public transport service has a planned daily schedule (service timetable). If this service serves $S_p = \{S_{p,1}, S_{p,2}, ..., S_{p,|S_p|}\}$ public transport stations and has a number of n_p trips during the day, where the trips are denoted as $K_p = \{K_{p,1}, K_{p,2}, ..., K_{p,|K_p|}\}$, then the planned daily schedule of service p is represented by a D: $|K_p| \times |S_p|$ dimensional matrix with integer elements denoting the planned arrival time of each trip $K_{p,j} \in K_p$ to each station $S_{p,i} \in S_p$.

Let $S_{p,i} \in S_p$ be the station of trip $K_{p,i} \equiv \pi_i \in K_p$ where one or more boardings from joint leisure activity passengers can occur and $S_{p,j} \in S_p$ be the alighting station of the joint leisure activity passenger/s which is the closest station to the location of the joint leisure activity Λ. The joint leisure activity passenger will be more tempted to use this public transport service if the alighting at the destination station occurs close to the time of the joint leisure activity. Given the joint leisure activity location Λ and the starting time of this activity t, a utility score for using the public transport service is introduced:

$$z = (t - t^w - D_{K_{p,i}, S_{p_j}})^2 \tag{5.1}$$

which is zero at the optimal case (alighting at station $S_{p,j}$ occurs at the same time with the starting time of the joint leisure activity t minus the walking time required from the station to the location of the activity denoted as t^w) and it progressively increases its value if trip $K_{p,i}$ arrives much earlier or later to that station.

During the rescheduling phase of this public transport service, the departure times of the daily trips are subject to change. Those changes are the decision variables ("unknows") of the rescheduling problem and are denoted by $x = \{x_{K_{p,1}}, x_{K_{p,2}}, ..., x_{K_{p,|K_p|}}\}$ which is a $|K_p|$-dimensional vector of the problem variables and represents how many minutes the departure time of each trip of public transport service p can deviate from the planned one. Vector x is not allowed to take values from the Euclidean space $\mathbb{R}^{|K_p|}$ but only from a discrete set of values ($q = \{-5, -4, ..., 0, ..., +4, +5\}$minutes) because the re-scheduled timeplan should be expressed in minutes (decimal values are not allowed to planned schedules). Therefore, the utility score for the joint leisure activity participant towards the public transport service can be expressed by a scalar function $z : \mathbb{N}^{|K_p|} \rightarrow \mathbb{R}$:

$$z(x_{K_{p,i}}) = (t - t^w - (D_{K_{p,i},S_{p_j}} + x_{K_{p,i}}))^2 \tag{5.2}$$

Realizing that an entire re-scheduling of all daily trips is unnecessary and will create many changes without reason, variables $x_{K_{p,\phi}} \in x$ are allowed to change only if the planned arrival time $D_{K_{p,\phi},S_{p_j}}$ of the trip $K_{p,\phi}$ at station S_{p_j} is within the range of ±60minutes from the joint activity starting time minus the expected walking time to the activity location:

$$x_{K_{p,\phi}} : \begin{cases} \in q = \{-5, -4, ..., 0, ..., 4, 5\} & \text{if } |D_{K_{p,\phi},S_{p_j}} - t - t^w| \leq 60\text{minutes} \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

One might reasonably claim that rescheduling the public transport service timeplan for adapting to the joint leisure activity demand might penalize other regular passengers of the public transport service who are not involved in leisure activities. In practical terms, it in not feasible to monitor the effect of the rescheduling phase by continuously surveying the passengers of the public transport service. However, one can ensure that i) the operations are better of, or at least not deteriorated, for all passengers of that public transport service at a system-wide level and ii) no passengers are over-penalized from those changes.

Joint leisure activity spatio-temporal demand variations have greater impact to the services of dense metropolitan areas due to the volume of leisure demand changes that justifies the public transport rescheduling efforts. Those services in metropolitan areas are high frequency services and they are regularity-based (their Key Performance Index (KPI) is the service-wide Excess Waiting Time (EWT) score instead of punctuality-based On-Time-Adherence (OTA) of operations to the planned schedule). Regularity-based services dictate that the daily trips of one public transport service should keep a certain, pre-defined time deviation (headway) when passing by each station of the service which should not be too small (bunching) or too high leading to excess waiting times for passengers that wait at the station for that service.

The operational performance of the public transport services are assessed through this EWT scheme at high-frequency services, and, in some cities, such as London and Singapore, the public transport operators receive monetary penalties or bonuses depending on the daily EWT score of each service. In many other cities, the EWT score of the public transport service indicates the performance of the service and the service operator can have his contracts terminated if the daily EWT scores of that services are below average.

The service-wide EWT of one public transport service is a linear function of the EWT scores observed at several stations of the transport service during different periods of the day (morning peak, afternoon peak, etc.). Those time periods have starting and ending times: $((T_1^s, T_1^e),(T_2^s, T_2^e),(T_o^s, T_o^e))$. By definition, the EWT at each station, $S_{p,j} \in S_p$ is calculated from the planned arrival times of consecutive trips for every time period $((T_g^s, T_g^e))$ and since it is a variance from the expected waiting time for passengers cannot take a negative value:

$$EWT_{S_{p,j}}^{(T_g^s,T_g^e)} = \max\left[0; \frac{\sum_{i=2}^{|K_p|-1} \theta_{K_{p,i},S_{p,j}}((D_{K_{p,i},S_{p,j}} + x_{K_{p,i}}) - (D_{K_{p,i-1},S_{p,j}} + x_{K_{p,i-1}}))^2}{2\sum_{i=2}^{|K_p|-1} \theta_{K_{p,i},S_{p,j}}((D_{K_{p,i},S_{p,j}} + x_{K_{p,i}}) - (D_{K_{p,i-1},S_{p,j}} + x_{K_{p,i-1}}))} - \frac{\sum_{i=2}^{|K_p|-1} \theta_{K_{p,i},S_{p,j}}((D_{K_{p,i},S_{p,j}} + x_{K_{p,i}}) - (D_{K_{p,i-1},S_{p,j}} + x_{K_{p,i-1}}))}{2\sum_{i=2}^{|K_p|-1} \theta_{K_{p,i},S_{p,j}}}\right] \tag{5.4}$$

where

$$\theta_{K_{p,i},S_{p,j}} : \begin{cases} 1 & \text{if } D_{K_{p,i},S_{p,j}} + x_{K_{p,i}} \wedge D_{K_{p,i-1},S_{p,j}} + x_{K_{p,i-1}} \in (T_g^s, T_g^e) \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

The EWT score at station $S_{p,j}$ for public transport service p is then derived by aggregating the EWT scores from different time periods of the day:

$$EWT_{S_{p,j}} = \frac{\sum_{g=1}^{o} EWT_{S_{p,j}}^{(T_g^s,T_g^e)}}{\sum_{g=1}^{o} 1} \tag{5.6}$$

In a similar fashion, the service-level EWT which is the main KPI for every regularity-based services is then derived by adding the EWTs computed at different stations (in some cases some stations might have higher weights than others):

$$EWT_p = \frac{\sum_{j=1}^{|S_p|} EWT_{S_{p,j}}}{\sum_{j=1}^{|S_p|} 1} \tag{5.7}$$

For satisfying condition (i): "the operations are better of, or at least not deteriorated, at the service-level after the rescheduling of the service", the first objective is to reduce the EWT score at the service level or at least ensure that it is not increased. In addition, for not increasing operational costs, it is not allowed the insertion of new, additional transport modes at each re-scheduled public transport service and the number of planned trips is kept the same as before (no inclusion of additional trips). Changing however the departure times of some trips might lead to operational cost changes such as fuel consumption due to traffic condition changes but those costs are omitted from the analysis because they are of minor importance especially since our re-scheduled trips change departure times within the range of $q = \{-5, +5\}$minutes resulting to small changes (no morning trips are shifted to afternoon or from peak to off-peak conditions where the network traffic is expected to change significantly).

Finally, for satisfying condition (ii) and not allowing any regular passengers to get over-penalized because of the re-scheduling, the frequency constraint rule is adopted which is also used for the tactical schedule planning from public transport operators. This rule is the outcome of the tactical frequency setting phase and dictates the range of minutes where successive trips can depart from the departure station for covering adequately the passenger demand at every station of the planned service. The frequency range changes over different time periods of the day (morning peak, etc.) and if the time periods of the day are: $\{1, 2, ..., o\}$=$((T^s_1, T^e_1), (T^s_2, T^e_2), (T^s_o, T^e_o))$, then for each time period $g \in \{1, 2, ..., o\}$ there is a minimal frequency value $l(g)$ and a maximal frequency value $h(g)$ which should not be violated by any couple of successive trips within that time period. This rule indirectly ensures that any re-scheduled timeplan that does not violate the frequency range for every time period of the day will cover the regular passenger demand in a satisfactory manner and no regular passengers of the service will be over-penalized.

For instance, if one trip $K_{p,\phi}$ is subject to re-scheduling and belongs to the time period g, then the departure time deviation $x_{K_{p,\phi}}$ should be such that:

$$l(g) \leq (D_{K_{p,\phi}, S_{p,1}} + x_{K_{p,\phi}}) - (D_{K_{p,\phi-1}, S_{p,1}} + x_{K_{p,\phi-1}}) \leq h(g) \tag{5.8}$$

While re-scheduling the planned public transport service p that contains a joint leisure activity trip $K_{p,i} \in K_p$, the objective is to: (a) minimize the utility score of joint leisure activity passengers $(t - t^w - (D_{K_{p,i}, S_{p_j}} + x_{K_{p,i}}))^2$, (b) disturb only the planned trips which are at the close vicinity of trip $K_{p,i} \in K_p$: $x_{K_{p,\phi}} = 0$ if $|D_{K_{p,\phi}, S_{p_j}} - t - t^w| >$ 60minutes, (c) minimize the service-wide EWT score, (d) do not violate the planned frequency range of successive trips derived from the tactical planning phase, (e) ensure that the new EWT will at least not be worse than the service-level EWT of the original timeplan. It is important to note here that points (c) and (d) are also the objectives for the public transport authorities when they plan their daily timeplans (timeschedules). If the service-level EWT of service p before the re-scheduling was EWT_p^{before}, then the above objectives result in the following optimization program written in the standard form:

$$\begin{aligned}
\min_{x} \quad & f(x) = (t - t^w - (D_{K_{p,i},S_{p_j}} + x_{K_{p,i}}))^2 + EWT_p(x) \\
\text{s.t.} \quad & EWT_p^{before} - EWT_p(x) \geq 0 \\
& (D_{K_{p,\phi},S_{p,1}} + x_{K_{p,\phi}}) - (D_{K_{p,\phi-1},S_{p,1}} + x_{K_{p,\phi-1}}) - l(g) \geq 0 \\
& \quad ,\forall D_{K_{p,\phi},S_{p,1}} + x_{K_{p,\phi}} \in (T_g^s, T_g^e), \forall g \in \{0,1,2,...,o\} \\
& h(g) - (D_{K_{p,\phi},S_{p,1}} + x_{K_{p,\phi}}) + (D_{K_{p,\phi-1},S_{p,1}} + x_{K_{p,\phi-1}}) \geq 0 \\
& \quad ,\forall D_{K_{p,\phi},S_{p,1}} + x_{K_{p,\phi}} \in (T_g^s, T_g^e), \forall g \in \{0,1,2,...,o\} \\
& x \in q = \{-5,-4,...,0,...,4,5\}\text{minutes} \\
& x_{K_{p,\phi}} = 0, \forall \phi : |D_{K_{p,\phi},S_{p_j}} - t - t^w| > 60\text{minutes}
\end{aligned} \tag{5.9}$$

In this optimization problem, the objective function is nonlinear and the service-level EWT constraint is also non-linear. This problem is combinatorial and includes also a series of linear inequality constraints for satisfying the allowed frequency ranges. Item (a) is covered by the objective function. Item (c) is also covered by the objective function whereas item (e) is covered from the nonlinear EWT constraint. Item (d) is covered by the series of linear inequality constraints and item (b) is covered by setting $x_{K_{p,\phi}} = 0, \forall \phi : |D_{K_{p,\phi},S_{p_j}} - t - t^w| > 60$minutes. The computational complexity of this problem is exponential $O(|q|^\rho)$ where ρ is equal to the number of re-scheduled trips (i.e., trips for which $|D_{K_{p,\phi},S_{p_j}} - t - t^w| \leq 60$minutes); therefore, this problem is computational intractable and an exact solution can be computed only at small-scale scenarios. For instance, if within the 2-hour time period of interest there are 20 planned trips from this public transport service a minimum number of 6.727E+20 computations is required. In addition, if there is not one, but P public transport services that require rescheduling for adjusting to the leisure activity demand, then the computational complexity becomes $O(P|q|^\rho)$.

Finally, it is evident that if within the time range $D_{K_{p,i},S_{p,j}} \pm 60$minutes there is another trip $K_{p,m}$ that belongs to the same public transport service and can serve another joint leisure activity demand with $S_{p,m}$ the closest station to that activity location and t_1, t_1^w the starting time and the walking distance from station $S_{p,m}$ to the location of that activity respectively, then the re-scheduling problem of this public transport service is transformed to:

$$
\begin{aligned}
\min_{x} \quad & f(x) = (t - t^{w} - (D_{K_{p,i},S_{p_j}} + x_{K_{p,i}}))^2 + (t_1 - t_1^{w} - (D_{K_{p,m},S_{pm}} + x_{K_{p,m}}))^2 + EWT_p(x) \\
\text{s.t.} \quad & EWT_p^{before} - EWT_p(x) \geq 0 \\
& (D_{K_{p,\phi},S_{p,1}} + x_{K_{p,\phi}}) - (D_{K_{p,\phi-1},S_{p,1}} + x_{K_{p,\phi-1}}) - l(g) \geq 0, \\
& \quad \forall g \in \{0,1,2,...,o\}, \forall D_{K_{p,\phi},S_{p,1}} + x_{K_{p,\phi}} \in (T_g^s, T_g^e) \\
& h(g) - (D_{K_{p,\phi},S_{p,1}} + x_{K_{p,\phi}}) + (D_{K_{p,\phi-1},S_{p,1}} + x_{K_{p,\phi-1}}) \geq 0 \\
& \quad \forall g \in \{0,1,2,...,o\}, \forall D_{K_{p,\phi},S_{p,1}} + x_{K_{p,\phi}} \in (T_g^s, T_g^e) \\
& x \in q = \{-5,-4,...,0,...,4,5\}\text{minutes} \\
& x_{K_{p,\phi}} = 0, \forall \phi : |D_{K_{p,\phi},S_{p_j}} - t - t^{w}| > 60\text{minutes} \cup |D_{K_{p,\phi},S_{pm}} - t_1 - t_1^{w}| > 60\text{minutes}
\end{aligned}
\tag{5.10}
$$

5.2. Demand responsive public transportation rescheduling with evolutionary optimization

The proposed heuristic search method is a heuristic optimization method which attempts to explore intelligently the solution space and converge to the *global* minimum of the multivariate scalar function f. Note though that heuristics do not guarantee a convergence towards a globally optimal solution. The heuristic method is composed of the following features: i) formulate a penalty function $p(x)$ for including the constraints to the objective function and penalize the objective function score if any constraint is violated leading to the formation of an unconstrained optimization problem where $p(x) = f(x)$ when all constraints are satisfied ii) generate two random parents and proceed to next generations with the formulation of an iterative Genetic Algorithm (GA).

This global minimization method is heuristic and it is not possible to determine if the true global minimum has actually been found. Instead, as a consistency check, the algorithm can be run from a number of different random starting points to increase the chances that one of our solutions is close to the global minimum.

For notation simplification, let assume that the problem constraints (the service-level EWT score nonlinear inequality constraint and the frequency range linear inequality constraints) are denoted by $c_i(x) \geq 0, \forall i$ where i represents every different constraint. For instance $c_1(x) = EWT_p^{before} - EWT_p(x) \geq 0$. Then, the re-scheduling problem of Eq.5.9

for service p can be expressed in a simplified manner as:

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & c_1(x) \geq 0 \\ \forall g \in \{0,1,2,...,o\} & c_2(x), c_3(x), ..., c_{1+m1}(x) \geq 0, \forall D_{K_{p,\phi},S_{p,1}} + x_{K_{p,\phi}} \in (T^s_g, T^e_g) \\ \forall g \in \{0,1,2,...,o\} & c_{m1+2}(x), c_{m1+3}(x), ..., c_{1+m1+m2}(x) \geq 0, \forall D_{K_{p,\phi},S_{p,1}} + x_{K_{p,\phi}} \in (T^s_g, T^e_g) \\ & x \in q = \{-5,-4,...,0,...,4,5\}\text{minutes} \\ & x_{K_{p,\phi}} = 0, \forall \phi : |D_{K_{p,\phi},S_{p_j}} - t - t^w| > 60\text{minutes} \end{array} \tag{5.11}$$

Introducing a penalty function, $p(x)$, can transform the above constrained optimization problem to an unconstrained one:

$$\begin{array}{ll} \text{minimize} & p(x) = f(x) + \sum_{i=1}^{1+m1+m2} (\min[-c_i(x), 0])^2 \\ & x \in q = \{-5,-4,...,0,...,4,5\}\text{minutes} \\ & x_{K_{p,\phi}} = 0, \forall \phi : |D_{K_{p,\phi},S_{p_j}} - t - t^w| > 60\text{minutes} \end{array} \tag{5.12}$$

The expression $(min[-c_i(x), 0])^2$ is the main addition to the penalty function and dictates that if a constraint $c_i(x)$ is negative, then $\min[-c_i(x), 0] = -c_i(x)$ and the constraint is violated for the current set of variables x; therefore, the objective function is penalized by the term $(min[-c_i(x), 0] = -c_i(x))^2$. Otherwise, if $c_i(x) \geq 0$ this constraint is satisfied for the current set of variables x and is not penalizing the objective function since $\min[-c_i(x), 0] = 0$ and $(min[-c_i(x), 0] = 0)^2 = 0$.

The heuristic search has a dual scope. Starting from an initial solution guess x_0 where $p(x_0) \geq f(x_0)$ it should explore the solution space until at some iteration k there is $p(x_k) = f(x_k)$. At that iteration it is ensured that all constraints are satisfied. However, the search should continue to reduce further the score of $p(x)$ until a convergence test is satisfied at some solution x_* which is also the approximate solution to the search of the global optimum.

At the initial stage of the heuristic search, let assume that there is a public transport service p which can serve a joint leisure activity that starts close to station $S_{p,j}$ and a number of ρ trips belong to that service such that $|D_{K_{p,\phi},S_{p_j}} - t - t^w| \leq 60\text{minutes}, \forall \phi \in \{1,2,...,\rho\}$. Then, one can re-schedule the departure times of those ρ trips by selecting different values from the discrete set $q = \{-5,-4,...,0,...,4,5\}$minutes.

At the first stage two random sets (parents) of problem variables are generated. Let the first random set (parent 1) be denoted as Pa. Pa is a ρ-dimensional vector where each element of the vector contains a random value from the set q: $Pa = \{Pa_1, Pa_2, ..., Pa_\rho\}$ and $Pa_1, ..., Pa_\rho$ take random values from the set q. In addition, the second random set (parent 2) has exactly the same characteristics with Pa, namely, $Pb = \{Pb1, Pb2, ..., Pb\rho\}$ and $Pb_1, ..., Pb_\rho$ take random values from the set q. Finally, another random set (child) denoted as Ch is initialized for which also $Ch = \{Ch_1, Ch_2, ..., Ch_\rho\}$ and $Ch_1, Ch_2, ..., Ch_\rho$ take random values from the set q.

The heuristic search exploration is iterative and at the first stage the penalty function performance of the parent 1 and the parent 2 ($p(Pa)$, $p(Pb)$) is computed by selecting the penalty function as the fitness function. The parent with the highest penalty function score is the 'weak' parent and is the selected parent for replacement at future iterations. At this stage, a sequential crossover phase is triggered starting from the first element of both parents (Pa_1 and Pb_1) and choosing randomly one of those two values to replace the child value Ch_1. If Pa_1 is randomly chosen, then the child vector becomes $Ch = \{Pa_1, Ch_2, ..., Ch_\rho\}$. In addition, mutation is allowed by introducing a mutation parameter $m_u \leftarrow 1$. Then, a number a is randomly chosen from the integer set $\{0, 1, ..., 9, 10\}$ and if $a \leq m_u$, then the 1st child vector element is replaced again by a random value ra from the set q: $Ch = \{ra, Ch_2, ..., Ch_\rho\}$. In other words, during the mutation phase there is a 10% probability to change the randomly chosen element Pa_1 from the crossover stage with another random value ra from the discrete set q.

The fitness of the new child vector is then assessed by computing the penalty function for $p(Ch)$ and (i) if $p(Ch) < p(Pa)$ and $p(Ch) < p(Pb)$, then if $p(Pa) \leq p(Pb)$ the Pb is replaced with Ch: $Pb \leftarrow Ch$ (elitism), (ii) if $p(Ch) < p(Pa)$ and $p(Ch) > P(b)$ then Pa is replaced with Ch: $Pa \leftarrow Ch$ (the same holds also for the opposite, i.e., $p(Ch) < p(Pb)$ and $p(Ch) > P(a)$), (iii) if $p(Ch) \geq p(Pa)$ and $p(Ch) \geq p(Pb)$, then the child has inferior performance and no parent is replaced. In addition, the selected child value at this sequence is dropped. For instance, if $Ch = \{ra, Ch_2, ..., Ch_\rho\}$ after the crossover and mutation phase, then Ch drops the change of this sequence and becomes again $Ch = \{Ch_1, Ch_2, ..., Ch_\rho\}$.

The sequential crossover continues then with the second elements of parent 1 and parent 2: (Pa_2, Pb_2) and the same crossover/mutation/elitism replacement procedure continues for all trips $\{1, 2, ..., \rho\}$. When the crossover/mutation/elitism replacement procedure for the departure time deviation of trips ρ is reached, then the procedure re-starts again from the beginning performing another sequential GA search. The first objective is to run this artificial evolutionary approach repeatedly until a child which has a penalty function score equal to the objective function $p(Ch) = f(Ch)$ is found (something that guarantees that all constraints are met for the departure time deviations denoted by that child).

At some point, parent sets Pa, Pb might have become too similar after their continuous replacement with better-performing children. Then, the only way to generate a new child with different characteristics is via mutation but this might take too long if the mutation rate $m_u = 1$. For this reason, if at some point $b \in \{1, 2, ..., \rho\}$, $Pa_b = Pb_b$ the mutation rate probability is increased to $m_u = 7$. The same applies when for too many iterations (more than 500) there is no new child generation with better performance than its parents. When this new child is finally found, the mutation rate is restored back to $m_u = 1$ for avoiding a totally random generation of new children.

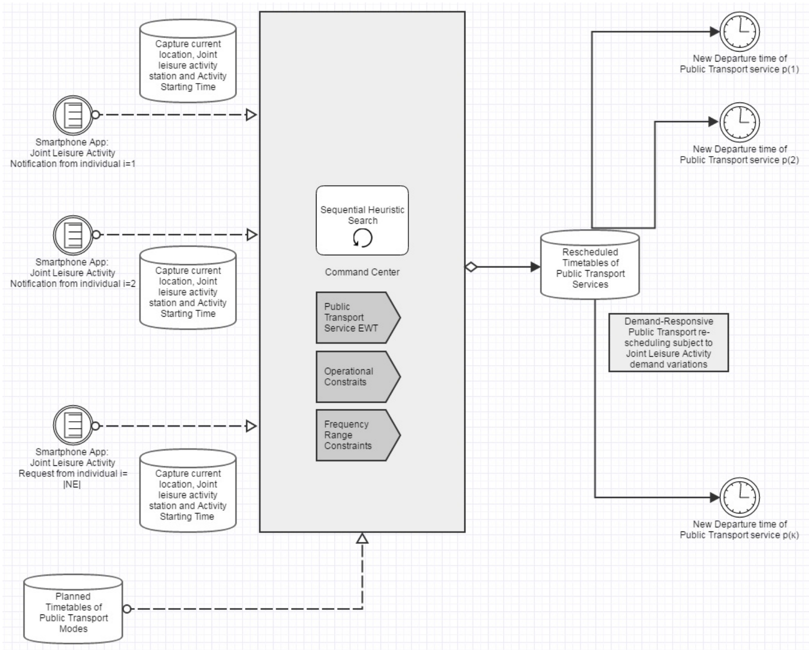


FIGURE 5.1.: Generic representation of the demand-responsive public transportation subject to operational KPIs and ridership maximization for joint leisure activities

The final stage is the heuristic search termination criterion which is controlled by two elements. The first is the maximum allowed number of iterations max_it and the second is the convergence rate. For the convergence rate criterion, if for many iterations there is not a new child that performs better than its parents even if the mutation rate is set to $m_u = 7$ and the previous children that replaced their parents had too close performance while also $p(Pa) = f(Pa)$ and $p(Pb) = f(Pb)$, one can assume that a minimum point (bottom of a valley) is reached (that can be any of the local optimums or the global optimum). Since there is no mechanism to justify that the minimum point is the global optimum, the algorithm can be terminated. The heuristic search method is summarized in the following algorithm:

The sequential heuristic search has computational complexity $O(max\_it \times \rho)$ and in the case of L public transport services that require rescheduling, each service has a unique computational complexity depending on the number of trips inside that service that need to be rescheduled. With the sequential heuristic search the time complexity is reduced from exponential to polynomial and a solution search is feasible even for large values of ρ; however, the proposed solution converges to an approximate global optimal which cannot be guaranteed that is the real global optimal.

The re-scheduled departure times that improve the service-wide EWT and adjust the service to the joint leisure activity demand should be finally adopted by the Demand Responsive public transport service. For this, a seamless operation Command Center receives information from Smartphone Apps for deriving the current location of individuals on the transport network and the location and arrival times of their planned joint leisure activities. Later, the heuristic sequential search solution method is applied for computing the desired departure times of the public transport modes (Fig.5.1).

Algorithm 3: Sequential Heuristic Search($p(x)$, $f(x)$)

Set $m_u \leftarrow 1$, $q = \{-5, -4, ..., 0, ..., 4, 5\}$ and iteration←0;
With an *initial random guess* choose a parent 1 set ($Pa = \{Pa_1, Pa_2, ..., Pa_\rho\}$) and compute $p(Pa)$;
With an *initial random guess* choose a parent 2 set ($Pb = \{Pb_1, Pb_2, ..., Pb_\rho\}$) and compute $p(Pb)$;
With an *initial random guess* choose a child set ($Ch = \{Ch_1, Ch_2, ..., Ch_\rho\}$);
while *iterations* $\leq$ *max_it (1st termination criterion)* **do**
    iteration←iteration+1;
    for *i in range* $\{1, 2, ..., \rho\}$ **do**
        Crossover: Set $c \leftarrow$ random choice between Pa_i and Pb_i;
        Mutation: if $a \leq m_u$ where a is a random choice number from the range $\{0, 1, ..., 9, 10\}$, then $c \leftarrow$ random choice value from the set q;
        Replace the i^{th} element of the Ch vector with c;
        Compute $p(Ch)$;
        if $p(Ch) < p(Pa)$ *or* $p(Ch) < p(Pb)$ **then**
            Elitism: Replace the parent with the highest penalty function with Ch;
            Set $m_u \leftarrow 1$;
            Set $failed\_changes\_counter \leftarrow 0$;
        else
            Undo the replacement of the i^{th} element of the Ch vector with c (set it back to its previous Ch_i value);
            Set $failed\_changes\_counter \leftarrow failed\_changes\_counter + 1$;
        end
        if $p(Pa) = f(Pa), p(Pb) = f(Pb), p(Pa) \simeq p(pB)$ **then**
            If for many generations no child had better; performance than its parents: ($failed\_changes\_counter > 10000$), assume convergence and terminate (2nd termination criterion);
        end
    end
end
return optimal solution x^* where $x^* = Pa$ if $p(Pa) \leq p(Pb)$ and $x^* = Pb$ if $p(Pb) \leq p(Pa)$;

5.3. Optimizing the public transport schedules

For the optimization, GTFS data from Sweden including the planned schedule of public transport modes for the period 13 February 2016 - 17 June 2016 is utilized. The data includes the files of Table 5.1.

For deriving the planned schedules of public transport modes, a library was developed in Python 2.7. The library processes .txt files and converts/stores them to an sql database. This facilitates data queries and enables web-based visualization (Fig.5.2) of the public transport operations with the use of `OpenStreetMap` (via `OpenLayers`, an open-source `JavaScript` library: `http://www.openlayers.org/api/OpenLayers.js`). The developed Python GTFS library: i) converts .txt files to sql database tables, ii) can query public transport routes from the database tables, iii) creates new files containing the planned trips for each route in ascending order (starting from the earlier morning trip to the latest night trip).

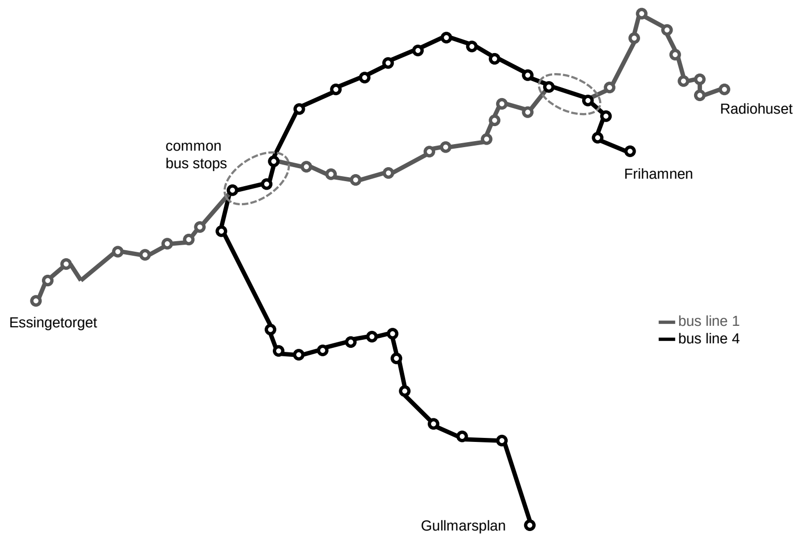


FIGURE 5.2.: Visualization after querying bus services 1 and 4 from Sweden GTFS data

After applying the Python GTFS library, the planned trips for every public transportation service are sorted and for each trip one can get the planned arrival time at every station. In particular, the study focus is on two bi-directional central bus lines (1 and 4) in Stockholm which can be seen as 4 independent services because every line direction has another EWT score and another set of constraints. Those are bus service 1 (bus line 1, direction 1 (Essingetorget to Stockholm Frihamnen)), bus service 2 (bus line 1, direction 2 (Stockholm Frihamnen to Essingetorget)), service 3 (bus line 4, direction 1 (Gullmarsplan to Radiohuset)) and service 4 (bus line 4, direction 2 (Radiohuset to Gullmarsplan)). In

Table 5.1.: GTFS data files and size (total size: 0.25GB)

File Name	Size	File Name	Size
agency.txt	5KB	stop_times.txt	242,000KB
calendar.txt	17KB	stops.txt	6,384KB
calendar_dates.txt	890KB	transfers.txt	1,020KB
routes.txt	218KB	trips.txt	11,628KB

Fig.6.4 the stations of services 1, 2, 3, 4 for both directions are presented together with the EWT at each station calculated from the planned arrival times according to the Eq.5.4.

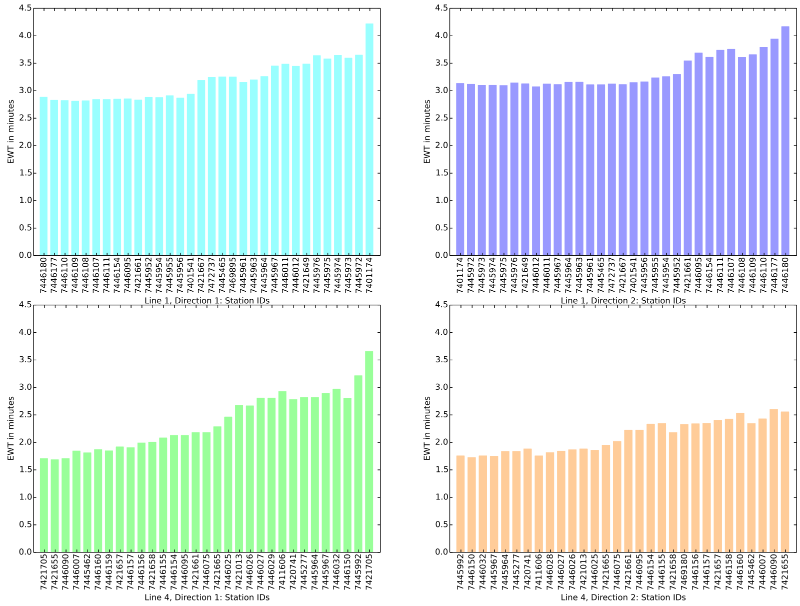


FIGURE 5.3.: EWT at every station and service-level EWT for every service according to the planned schedule of daily trips for bus lines 1 and 4 in both directions for the afternoon EWT phase (time period 2:10pm-7:30pm)

Apart from public transportation data, the data of travelers that are heading to joint leisure activities is also retrieved following the methods described in the previous chapters and utilizing social media data from Stockholm. For the targeted data mining of Twitter, a platform in Python 2.7. is utilized. From this data, the geo-location information from $N_E^* = 62$ individuals from Stockholm is collected. Those individuals were traveling from their current location to other locations of joint leisure activities. Those 62 individuals were selected according to the following criteria: i) they change geo-location outside working hours (from 15:30-19:30); ii) the location they are heading is not related to work/home, but it is a leisure activity location for them; iii) their final destination is close to one of the interchange stations of bus lines 1 and 4 (namely, bus stations 7446154, 7446095, 7421661, 7445964, 7445967 and 7445952); therefore, they have the option to use bus line 1 or bus line 4; iv) their travel origin is within walking distance from at least one bus station of lines 1 and 4; v) they arrive at the location of the joint leisure activity within a time variance of less than 25min.; therefore, it is assumed that they participate at the same activity. Criterion (v) is the strongest assumption since multiple joint leisure activities can occur nearby concurrently. However, by depending solely on Twitter data one cannot justify the potential splitting of a joint leisure activity into multiple concurrent ones with a high degree of confidence. The current locations and the final destinations of groups of individuals who are heading to the same leisure activity location are presented in Table 5.2.

Table 5.2.: Nearest stations to individuals who are performing a joint leisure activity

Individuals who are performing similar joint leisure activities	($N_E^* = 62$: individual trips from current station to Joint Leisure Activity Station)	Trip Destination Bus Station (Joint Leisure Activity Station)	Event Time
Trip Departure Bus Station	7445462, 7446157, 7421658, 7469180, 7401541	7446154	17:00
Trip Departure Bus Station	7445961, 7401541, 7445964, 7421665, 7421657, 7446075, 7446028, 7421661	7445967	16:30
Trip Departure Bus Station	7446095, 7446026, 7446025, 7445954, 7421649, 7421661, 7446177, 7445976, 7445974, 7446026, 7445955	7445964	16:00
Trip Departure Bus Station	7446108, 7446075, 7445992, 7421705	7445964	19:30
Trip Departure Bus Station	7446108, 7446177, 7446160, 7421657, 7446011, 7469180, 7446156, 7446109, 7446177	7446095	17:00
Trip Departure Bus Station	7446007, 7445972, 7445465, 7445952, 7446150	7446095	18:30
Trip Departure Bus Station	7446090, 7445964, 7421661, 7446154, 7445952, 7446160, 7446157	7421661	15:30
Trip Departure Bus Station	7446029, 7421705, 7420741, 7469180, 7421705, 7421665, 7445992	7421661	18:00
Trip Departure Bus Station	7421667, 7445976, 7446095, 7446095, 7445967, 7421655	7445952	17:00

From the above, there are $\Lambda = 9$ different joint leisure activity locations and each one has the following number of participants who are heading there from their previous location: $N_E^* = 62$, where $N_E^* = N_E(\Lambda_1) + N_E(\Lambda_2) + ... + N_E(\Lambda_9)$ and $N_E(\Lambda_1) = 5$, $N_E(\Lambda_2) = 8$, $N_E(\Lambda_3) = 11$, $N_E(\Lambda_4) = 4$, $N_E(\Lambda_5) = 9$, $N_E(\Lambda_6) = 5$, $N_E(\Lambda_7) = 7$, $N_E(\Lambda_8) = 7$, and $N_E(\Lambda_9) = 6$.

On another note, Python GTFS library was utilized to query the database tables and select the number of public transport trips, π, that can serve those joint leisure activities. First, all planned trips from 14:10(850min.) - 19:30(1170min.) of bus lines 1, 4 on both directions together with their planned arrival times at every bus station are presented in Fig. 6.5. Then, those trips, π, that can be used from the $N_E^* = 62$ individuals for arriving at the locations of joint leisure activities are selected (refer to Table 5.3).

Table 5.3.: Trips IDs of π=97 trips that can be used from individuals for arriving to the locations of joint leisure activities

Line 1, Direction 1: 90656299, 90656481, 90656480, 90656293, 90656392, 90656601, 90656605, 90656408, 90656408, 90656613, 90656613, 90656607, 90656607, 90656409, 90656409, 90656615, 90656337, 90656725, 90656801, 90656271, 90656794
Line 1, Direction 2: 90655951, 90655947, 90655891, 90655953, 90655963, 90655963, 90655963, 90655886, 90655886, 90655886, 90655964, 90655824, 90655769, 90655903, 90655767, 90655938, 90655927, 90656128, 90656164
Line 4, Direction 1: 90661185, 90661816, 90661826, 90661821, 90660993, 90661189, 90661798, 90661808, 90661687, 90661832, 90661691, 90661835, 90661835, 90661519, 90661519, 90661519, 90661813, 90661693, 90661693, 90661428, 90660717, 90661426, 90660755, 90660756, 90661494, 90661651, 90661325
Line 4, Direction 2: 90661590, 90661059, 90660931, 90660962, 90661597, 90661585, 90661614, 90661616, 90661616, 90661037, 90661037, 90661588, 90661588, 90660956, 90660956, 90660956, 90661586, 90661586, 90661586, 90660939, 90661591, 90660946, 90660905, 90661035, 90660908, 90660977, 90661069, 90660918, 90661127, 90660920

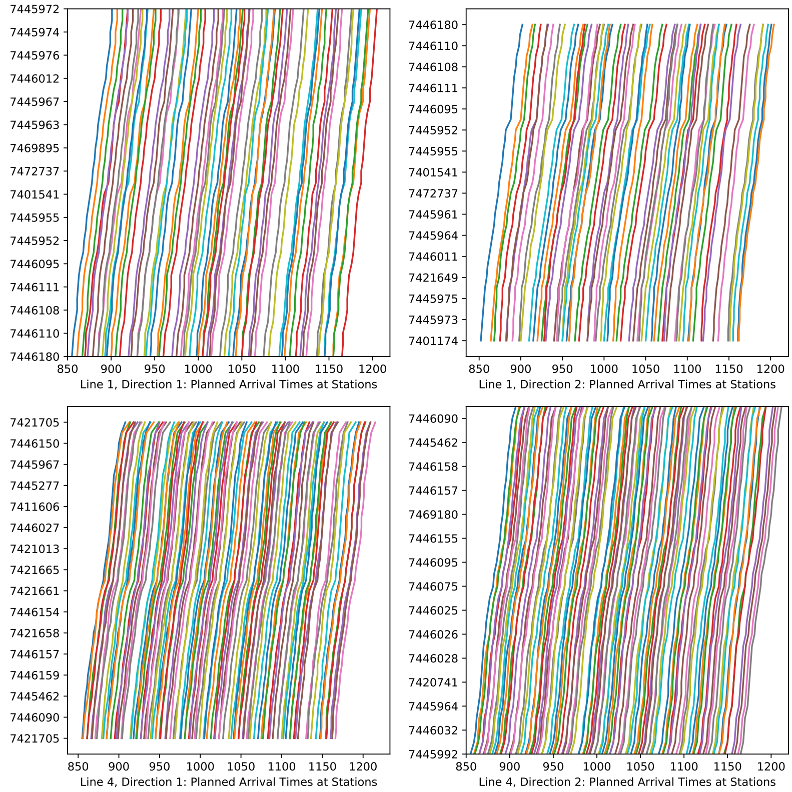


FIGURE 5.4.: Planned arrival times of trips at every station for bus services 1, 2, 3, 4 from 2:10pm (850min.) until 7:30pm (1170min.)

Those trips are planned to arrive at the joint leisure activity locations $\{\Lambda_1, \Lambda_2, ..., \Lambda_{|\Lambda|}\}$ at times which are close to the starting time of the joint leisure activity. In Fig.5.5 the planned arrival times of trips $\{\pi_1, \pi_2, ..., \pi_{\pi=97}\}$ at every joint leisure activity station and their deviation from the starting time of the joint leisure activities (which should be as close as possible to zero after the re-scheduling phase) are presented.

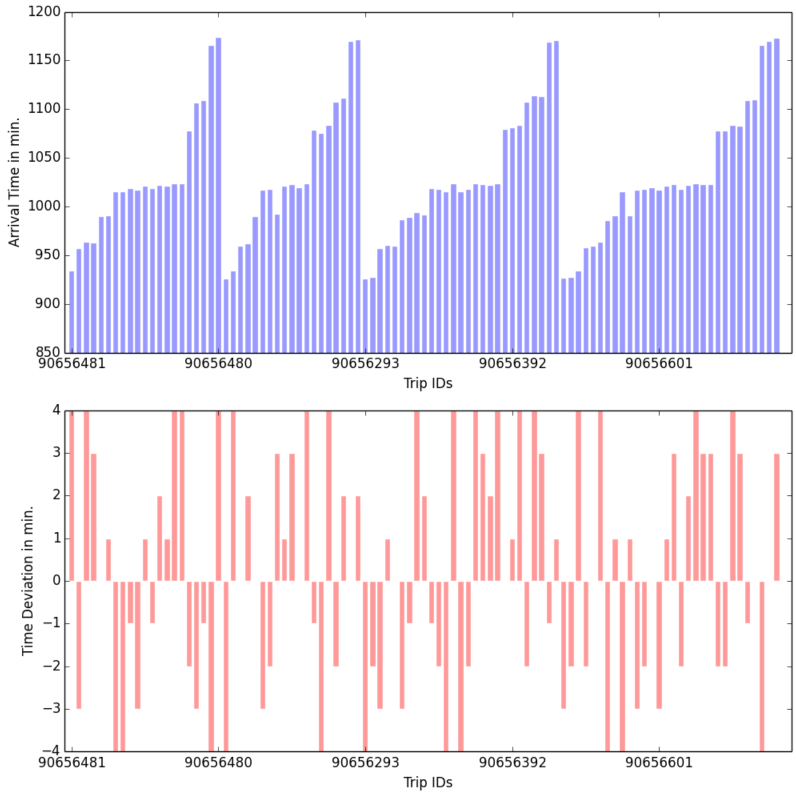


FIGURE 5.5.: Planned arrival times of trips $\{\pi_1, \pi_2, ..., \pi_{\pi=97}\}$at joint leisure activity stations $\{\Lambda_1, \Lambda_2, ..., \Lambda_{|\Lambda|}\}$ and their time deviation from the starting time of the activity in min. A positive deviation means that the bus arrived minutes after the activity started and a negative that arrived before

After optimizing the schedule for the bus services with the sequential heuristic search method, the new EWTs at every station and the new service-wide EWT are presented in Fig.5.6 in comparison with the planned schedule EWTs at the do-nothing scenario. While rescheduling each one of the public transport services that contain potential joint leisure activity trips $\{\pi_1, \pi_2, ..., \pi_{\pi=97}\}$ the objective is to: (a) minimize the deviation between the arrival time of each trip at the joint leisure activity location and the starting time of the activity (presented at Fig.5.5), (b) disturb only the planned trips which are at the close vicinity of trips $\{\pi_1, \pi_2, ..., \pi_{\pi=97}\}$ (less than 60minutes deviation), (c) minimize the service-wide EWT score, (d) do not violate the planned frequency range of successive trips derived from the tactical planning phase, (e) ensure that the new EWT will at least not be worse than the service-level EWT of the original timeplan.

		service-wide EWT	service-wide ASQ scores	Computational Time (sec)
Line 1 Direction 1	before	1.133273406	178	
	after	**0.75732**	**67**	129.44
Line 2 Direction 2	before	1.127213956	129	
	after	**0.904360219**	**44**	112.981
Line 4 Direction 1	before	1.05272815	210	
	after	**0.94029**	**97**	263.067
Line 4 Direction 2	before	0.840175194	216	
	after	**0.704939879**	**105**	199.558

Table 5.4.: Re-scheduling Summary Results for every service before and after the Sequential Heuristic Search Optimization

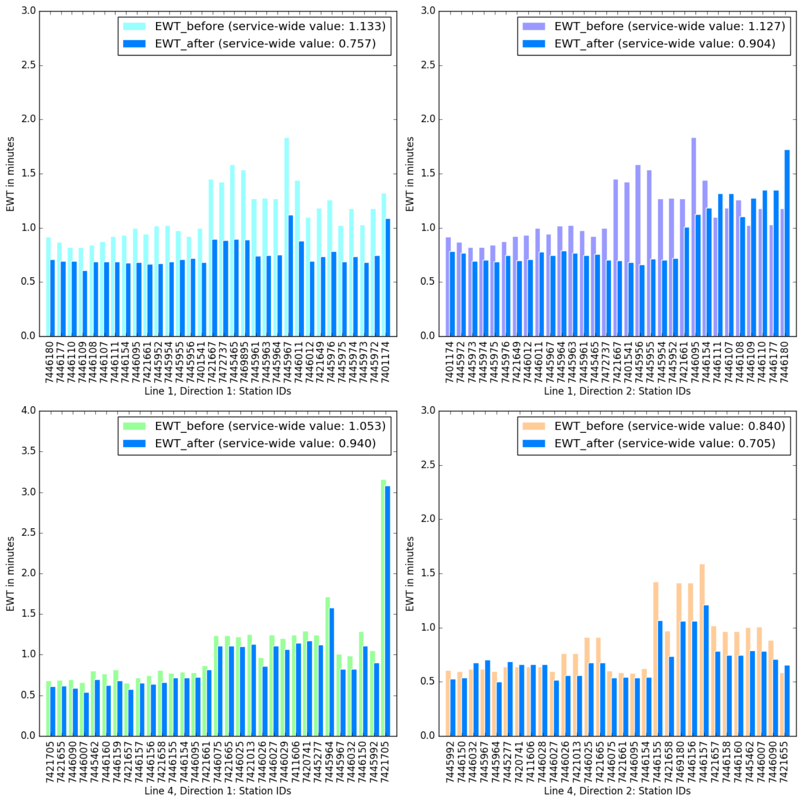


FIGURE 5.6.: EWT at every station and service-level EWT for every service before and after the heuristic search re-scheduling for the afternoon EWT phase (time period 2:10pm-7:30pm)

In addition, the service-wide EWT score before and after the optimization together with the value of the penalty functions at the convergence state are presented in Table 5.4. In that table it is also shown how the arrival time of trips π at the joint leisure activity locations has been adjusted to the joint leisure activity starting times by presenting the time deviation scores (computed from Eq.5.2) before and after the rescheduling. Those scores are the aggregated values of the squared difference of the trips' arrival times to joint leisure activity locations and the starting times of those activities represented by the abbreviation ASQ scores.

For demonstration purposes also, the re-scheduled arrival times of bus trips $\pi_i \in \pi$ which can be potential joint leisure activity trips at the location of those activities are presented. The starting time of those activities are also presented. The re-scheduled trips have closer arrival times to the starting times of the activities after optimizing them with the sequential heuristic search method. How closer are the re-scheduled trip arrival times to the activity starting times compared to the planned arrival times is also presented in Fig.5.7.

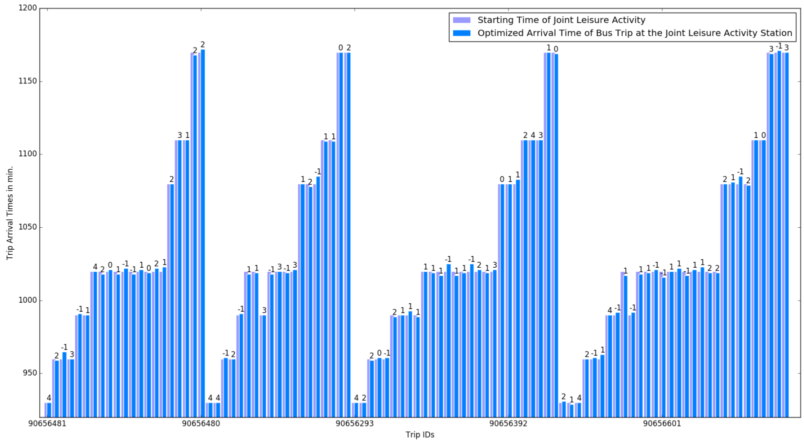


FIGURE 5.7.: Starting times of Joint Leisure activities and Re-scheduled arrival times of trips $\{\pi_1, \pi_2, ..., \pi_{\pi=97}\}$at joint leisure activity stations $\{\Lambda_1, \Lambda_2, ..., \Lambda_{|\Lambda|}\}$ in minutes. For every trip, the re-scheduling improvement in terms of adjustment closer to the starting activity time is also expressed in minutes (negative values indicate deterioration).

Finally, to show more detailed results on the computational costs, different scenarios with different numbers of public transport mode trips $\pi \leq 10$ are also presented in Fig.5.8, since a brute-force computation for $\pi = 97$ is not feasible due to the exponential time complexity.

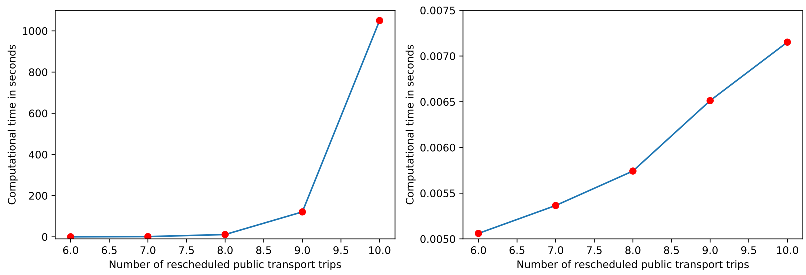


FIGURE 5.8.: Computing time for re-scheduling public transport operations related with joint leisure activities with brute force (left) and heuristic search (right). Only up to π=10 trips were able to be tested due to the exponential computational cost of brute-force

5.4. Discussion on the results

In a short summary, the public transport re-scheduling for covering joint leisure activity passenger demand subject to the no-deterioration of service quality and the adherence to a set of operational regulations was modeled and optimized. A sequential heuristic search algorithm was developed for this purpose in order to change the public transport schedules in near real time (in a matter of minutes) for adjusting to the arrival time needs of joint leisure activity passengers without deteriorating the KPIs of public transport operations.

In Fig.5.6 the re-scheduling changes on the bus line 1, 4 EWTs were presented demonstrating that the operational performance of bus services had an EWT improvement at a service-wide level for all services while only some stations from line 1, direction 2 and line 4, direction 2 had a slight EWT deterioration(up to 0.6min.). In addition, those under-performing stations did not affect significantly the level of service of bus operations. After re-scheduling, joint leisure activity trips from all services enjoyed new arrival times to the joint activity stations which were closer to the starting times of those activities. Finally, the computational cost of the proposed heuristic algorithm for public transportation re-scheduling was presented in Table 5.4 demonstrating that a convergence to the approximate global optimum requires from 2-6 minutes.

A straightforward expansion of this application is the utilization of the public transport mode optimization method for covering joint leisure activity trips at the entire public transport operations of a city (including bus, metro, trams and rail).

6. Journey planning optimization of Joint Leisure Activity Travelers under resource constraints

In previous chapters the main focus was on identifying users' mobility/activity patterns, optimizing joint leisure activities locations and times and re-scheduling public transport mode schedules for covering passenger demand related to those activities. After that, selecting the optimal route for reaching the joint leisure activity location is the last remaining step for optimizing joint leisure activities. Scientific research has shown that the lack of real-time information and/or the deficiency of communication mechanisms have prevented the optimal use of transport infrastructure. An attempt to give an idea of the degree of inefficiency was made in the experiment of [81]. In this experiment 188 routes between 91 Origin-Destination pairs were tested and only the 37% of the users chose the fastest path to their destinations due to lack of information.

The need for information exchange is served in transport networks by a range of communication technologies and Intelligent Transport Systems (ITS). Yet, this is an increasingly developing field and there is room for further improvements. In this thesis, the focus is mainly on mobile devices that contain also navigation capabilities. From now on, the term mobile device will be used to describe the new generation of smartphones with built-in Global Positioning System (GPS) receivers.

In general, mobile phones have many advantages both from the perspective of an individual user and the perspective of a network's manager. Considering the former case, mobile devices serve the need of practicality because not everyone wants a separate device for every function, even if that makes each one perform its specific task better. Moreover, mobile devices could be used for navigation purposes from every user (i.e., pedestrians, cyclists) and not only from motorists. As a result, mobile navigation opens new markets for different groups of customers, and is currently considered as the new competitor of in-vehicle satellite navigation. Until 2005, the navigation projects in mobile phones came a long way along the priority list, after higher-risk projects like mobile TV. However, the news from the global markets are promising and a survey on the use of navigation devices by GfK retail and technology (2010) stated that 53% of the respondents in Germany, 48% in UK and 39% in France articulated the belief that mobile phones with new navigation features would compete in-car satellite devices for market domination.

From a global perspective (i.e., system operator, traffic manager) the mobile navigation is also beneficial. Firstly, mobile devices can be used as data sensors by pinpointing the location of their user or providing information for other traffic parameters (i.e., average speed, velocity and more). This concept has already been used for in-vehicle navigation systems (probe vehicles), where the cost of the sensor was included in the price of the guidance system itself. The main difference though is that mobile devices are expected to saturate the market which means that they would be able to provide larger amounts of sensing data. Furthermore, more users will be equipped with navigation devices; hence, information about the real-time traffic conditions of the network, emergencies or accidents etc. would be disseminated to a larger audience.

Despite their advantages, mobile navigation systems suffer from several drawbacks. The two major disadvantages of mobile devices are their limited storage space and their low computing power. Nevertheless, the focus is on mobile navigation and how the comparative advantages of mobile navigation can be used for proposing an adequate mobile application for joint leisure activity journey planning. For this purpose, 2 different architectures are presented which have been used for route guidance systems and a set of requirements for the suggested mobile application is formed.

6.1. Route guidance systems for in-vehicle and mobile navigation devices

Regarding the provision of information to drivers, route guidance could be either prescriptive or descriptive. In the first case, it provides routes to the guided users, while in the second provides information only about the network state, letting the users interpret this information in their own terms.

The required data being used for navigation purposes may be consisted of network topology information, historical average conditions and real-time sensed data (either by stationary or mobile sensors). Corresponding to the nature of provided information, route guidance could be provided either just prior to departure or it could be continuously updated while the passenger is en-route. A proven way to broadcast live-traffic information to users is the Traffic Message Channel (TMC) of the Radio Data System (RDS), which makes use of coded traffic information transmitted with FM radio. However, considering the case of mobile navigation, another way for the dissemination of traffic information is the mobile Internet.

The location of the computational components categorize the route guidance into centralized or decentralized. In centralized route guidance exists only one central computational component which operates on data from the entire network and provides detailed, individual routes to its clients. On the other hand, in decentralized architectures, the route selection function is located on in-vehicle or mobile navigation devices. Additionally, in the autonomous form of the decentralized route guidance, all route guidance functionality is embedded in the navigation system, and this yields a static route guidance application. In decentralized dynamic architectures, link travel times are broadcasted to travelers to provide real-time estimates of network's condition and each navigation system computes its client's route.

Generally, decentralized systems are envisioned as reactive due to their low computational burden and their quick response to changing conditions (a good performance in the case of an emergency). Yet, they tend to lead to an all-or-nothing traffic assignment because different navigation systems which compute a route for the same Origin-Destination pair share the same disseminated information, and thus may propose the same path to their users. Finally, they tend to be poorer at prediction because they are not benefited from full data sharing. In contrary, in centralized systems, predictive algorithms which use historical and real time data together with models of network and user behavior are used to estimate future network conditions. Their drawbacks are their lack of reaction because they perform complex computational analysis and the influence of the network's conditions due to users' reaction to guidance. The latter is a complex problem being formed by the mutually dependent relationship between the prediction of network's conditions and the effects of the users' compliance to guidance (the effects of this problem could be mitigated by a sophisticated algorithm which predicts the future network state including the expected effect of guidance). Due to their drawbacks none of those route guidance methods has prevailed so far and new architectures, such as the Hybrid route guidance, have emerged to integrate those methods and use their strengths. For example, in a Hybrid route guidance system the predictive guidance is generated in a centralized layer and revised in a decentralized layer (for a comprehensive description see [82]).

[82]: Farver (2005), 'Hybrid vehicle-centric route guidance'

Considering the provision of route guidance, a transfer from the autonomous systems to the decentralized and finally to the centralized seems to be the meaningful way of evolution. The Simulator of Anticipatory Vehicle Network Traffic Flow (SAVaNT) developed by [83] was a first attempt to provide a predictive Decentralized Route Guidance where the central server disseminated information to its clients not only about the current link travel times, but also about the predicted link travel times within a predetermined time horizon. However, the anticipatory decentralized route guidance was superior to the non-predictive one if the market penetration of navigation systems (system's clients) was less than 30%.

DynaMIT is another software tool for centralized, predictive route guidance, first presented by [84] . This architecture enables a two-link communication between the users and the central server, and the external server itself provides route guidance to its clients with the intent of influencing their decisions. This concept is known as Consistent Anticipatory Route Guidance (CARG), where real-time traffic conditions are used to make predictions of the evolution of the network's traffic conditions within a predetermined time period. Hence, information provided to a user will reflect the conditions that are expected to prevail at network locations at the times when he will actually be there. One of the drawbacks in generating anticipatory route guidance is that the system under consideration is affected by the dissemination of information itself, thus the predictions on which the guidance was based may be invalidated. This problem was introduced by [85] while the first formulations of the CARG generation as a fixed point problem have been proposed by Bottom, [86] and developed by

[84]: Ben-Akiva et al. (1998), 'DynaMIT: a simulation-based system for traffic prediction'

[85]: Ben-Akiva et al. (1996), 'The impact of predictive information on guidance efficiency: An analytical approach'

[87].

6.2. Journey planning and guidance from mobile phones

Problem Description

One of the advantages of new generation mobile devices is their access to mobile internet. The online access could be used for multiple purposes, but first the focus is on map downloading. In this case, an external server encompasses digital map information which is distributed to mobile devices via mobile internet (e.g. Google maps). One of the benefits is that the user does not need to purchase any digital maps and then go through the process to install and retain them up-to-date. In addition, this service could be provided anywhere (provided that a mobile internet connection is established); hence, the user is not relying on local, pre-installed digital maps, but enjoys access to a digital map wherever he is. Finally, since the digital maps are not directly installed to the mobile device but uploaded via internet, there is no need for extra mobile storage space (the storage requirements are shifted from the mobile device to the external server).

Another matter of concern is the quality of the provided traffic data. A drive test by [88] (where the route guidance performance of two mobile and three in-vehicle devices where tested on field) demonstrated the importance of real-time traffic data for the selection of the fastest path. The device with the best performance was a TomTom HD 1000 (in-vehicle navigation device) with online access to TomTom's HD traffic centre, and this is good news for the idea of collecting live traffic data directly from mobile devices. However, live traffic data regarding the location of vehicles in the network and their speeds (which infers the travel time of network's links) may be enough for in-vehicle route guidance but is definitely not enough for the case of mobile navigation and the reason is that a significant proportion of mobile users does not hold a driver license or have a direct access to a vehicle. As a result, the aim of a mobile application should be to provide real-time multi-modal journey assistance for joint leisure activity travelers.

[88]: Sohr et al. (2011), 'Project drive test–comparison of five Personal Navigation Devices (PND)'

By the term multi-modal it is not described the provision of a fastest path solely by car or through a combination of different means of transport (which is already provided by Transport Direct, Google Navigation for Mobiles and more), but instead, the provision of the fastest route which integrates the use of car and other transport modes. Hence, the user will be informed about where to drive, where to park, what mode of transport to use, in which station should make a change between different transport modes and how to continue until the final destination. Of course, if the user does not have an access to a private vehicle he will be only informed about the fastest path via different modes of public transport. The real-time multi-modal journey assistance empowers the user to realize his options more accurately and it might also enhance inter-modal routing by suggesting private vehicle-public transport combined trips. However, in order to provide this functionality, a large amount of real-time and historical data should be provided to a central server via third party organizations. That data is summarized in Table 6.1

Table 6.1.: Data Requirements for the proposed mobile application

Road Data	Public Transport Data	Parking Data
Restrictions and Speed Limits	Public Transport Schedules	Parking Spots
Incident Information	Public Transport Routes / Stops	Parking Space Availability
Real Time Travel Time of Network Links	Real-Time Arrival and Departure Times	
Predicted Journey Times	Public Transport Incidents Information	

Considering the architecture of the mobile application, the anticipatory route guidance is adopted. This architecture requires information exchange between mobile devices and external servers, which will be called Mobile Device-External Server Partnership (M-ESP). In this architecture, the path-finding procedure is split in three distinct parts: a) the user's position is defined by the GPS receiver, b) up-to-date digital maps are instantly downloaded from the central server using a data plan tied to the device and c) the central server computes the optimal route and communicates it to its client.

The centralized architecture is preferred because it mitigates the weaknesses of mobile navigation. Firstly, the solution of the real-time multi-modal journey problem requires high computational power and this function cannot be supported by the typical 1000-2000MHz processor of a mobile device. Furthermore, the central server could provide disseminated information to road users with the intent of influencing their route choice decisions (this could yield a system-optimum traffic assignment, but this subject is not the purpose of this work). In addition, the central server computes the suggested route considering the future travel times at each link (concept of anticipatory route guidance). The main problem is that the system under consideration is affected by the dissemination of information itself, thus the predictions on which the guidance was based may be invalidated. A first idea is to use a hybrid architecture, where real-time information will be disseminated to the mobile device recursively and if a new fastest path is found (because the predictions of the server were not accurate) then a re-routing will be performed. However, even if hybrid architecture could be used to find the fastest path via a private vehicle, it is not recommended for a multimodal journey planning because in that case the mobile processor will not have the ability -in terms of computational power and according to what is known today- to provide a solution to the problem.

The computational power of the central server could be also used to provide more complex routes to its clients. For example, possible add-in functionality is the real-time multi-modal journey planning with respect to a number of constraints. The characteristics of M-ESP enable the development of a guidance system which not only offer travel assistance, but also serve other customers' preferences. The proposed architecture aims to go one step forward and enable the conversion of mobile devices from route assistants to personal travel guides. In order to achieve that, every mobile device should communicate with the external server and supply it with information about its owner's preferences (e.g. a user may not have a car or may prefer to walk less than 5 minutes). This will be served most properly if users' habits are stored in databases and the information is automatically retrieved in the future (i.e., every user would have a private account where his profile is stored or may be constantly retrieved from his/her profile settings that reside on the mobile device). Hence, the external server will be supplied with the required information and will suggest personal optimised routes based on each user's preferences. Of course, the main criterion for path-selection will still be the minimization of the total travel time (based on real-time traffic data and predictions about the links' travel time in the future) and the users' preferences (e.g., maximum fuel consumption, maximum walking time or maximum number of transfers between different modes) will provide threshold values which would lead to path rejection whenever they are violated.

Due to the creation of databases with observations about users' behavior, this architecture is expected to also benefit mass transit because a more qualitative traffic data would be collected from users of the application which could be used for transport modeling purposes. For example, regression based trip-end mode choice techniques in a macroscopic level or utility based techniques in a dis-aggregate level could be formed for the simulation of users' behavior and these techniques could be calibrated by comparing their results with data provided by the new architecture. Furthermore, one can test the validation of these techniques by studying different theoretical scenarios such as enforcing speed limits, diverting directions, offering a new bus service and explore possible users reactions to these measures. The service platform of the mobile application is presented in Figure 6.1.

The multi-modal journey problem under constraints

Following the presentation of the mobile application in the previous section, one should be cautioned that even if the required traffic data is provided (refer to Table 6.1), the real-time multi-modal journey problem under constraints requires high computational power and its solution is not a trivial task. In this section, a brief description of algorithms for route guidance will be provided in order to present the methods which can assist the solution of the routing problem.

The first route guidance methods were focused on the solution of the simple Point to Point Shortest Path Problem (PPSPP), and introduced in the late 1960s. More recently, a number of heuristic solutions have reduced that problem to one with an almost linear running time complexity. [89] , [90] and [91] have presented various algorithms for the solution of the PPSPP. Apart from the computation of the unconstrained fastest path, a new trend in route guidance considers also

[89]: Goldberg et al. (2006), 'Reach for a: Efficient point-to-point shortest path algorithms'

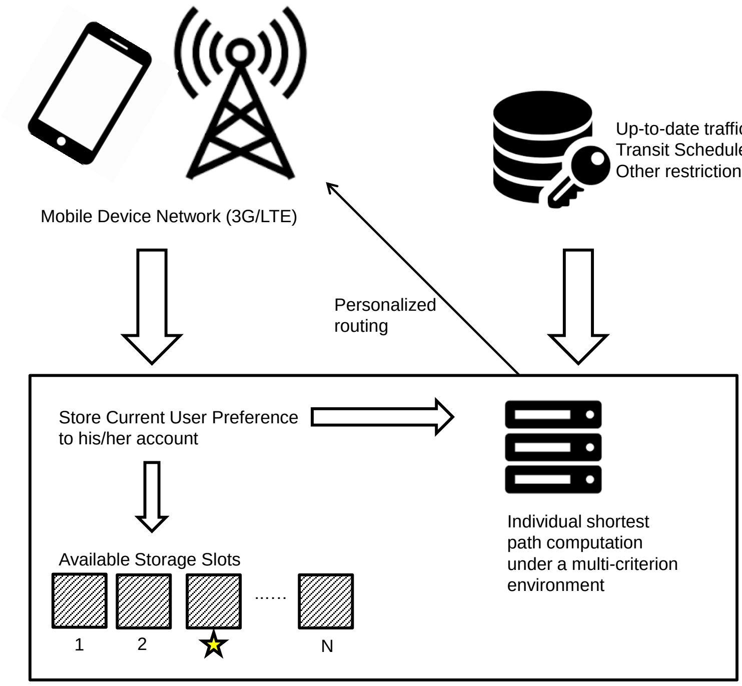


FIGURE 6.1.: M-ESP Service Platform

travel time uncertainty. [92] show that the link travel times in congested networks have a high degree of unreliability and in many cases the fastest path yields a worst solution compared to the more reliable one. In the paper of [93] , travel time reliability was considered as the objective function during the search for an efficient path and the total travel time or trip length were secondary objectives (they were regarded as additional constraints). In [94] a constrained risk-averse A* algorithm which solves the same problem was presented. It is generally assumed that travel time follows a log-normal distribution, and in the first step, the fastest path between two points is computed by using an appropriate PPSPP algorithm. Then, the links with unexpected (non-recurrent) delays are penalized. The procedure terminates when the computed route satisfies the constraints imposed. Following this method, non-recurrent delays are avoided and the proposed routes are not too circuitous, but it cannot be guaranteed that the new sub-optimal paths will be accepted by the users.

[93]: Rakha et al. (2006), 'Estimating path travel-time reliability'

The objective function of a route guidance problem still remains an open field in transport research and empirical studies by [95] , [96] and [97] indicated that unlike other problems -such as mode or destination choice- route choice is very difficult to be described because apart from the total travel time many other factors such as distance, reliability, personal preferences and tolls also affect the final decision. Regarding the proposed mobile application, the total travel time is considered as the objective function of the problem because the provided real-time traffic data mitigates the link travel times' unreliability problem. However, the other factors are not neglected and are considered as problem constraints.

[95]: Abdel-Aty et al. (1997), 'Using stated preference data for studying the effect of advanced traffic information on drivers' route choice'

Considering the computation of the fastest path in a multi-modal network -where transfers between private vehicles and public transport modes are allowed- the first step is to integrate the information about multiple modes in a single digital map. This could be done with the assistance of a Geographic Information System (GIS), where different layers contain information about the link travel times, the intersections and the stops and other attributes of each mode (e.g., layer 1 could represent the road network and its attributes, layer 2 the rail network, layer 3 the pedestrian network etc.). Following this idea, the mobile application and its routing guidance function will be based on a GIS system where each model will be represented by a different layer (for a more comprehensive description see [98] and [99]).

The second step is to develop an algorithms for the solution of the multi-modal travel planning problem with an acceptable running time complexity. The work of [100] , [101], [102] and [103] provide algorithms for the computation of the fastest path in a multi-modal network. Most algorithms are based on a label correcting approach that updates some labels associated with the graph nodes, where the transport network and the corresponding data are modeled as a multi-label graph. The objective function of the problem is the minimization of the total travel time from the origin to the final destination and if two paths have the same performance in terms of travel times the one with the less number of transfers is selected. In most cases, the departure times, the transport modes available in each node and the travel times on each link of the networks are considered as fixed. In multi-modal routing, one deals with the additional constraint of departure time from each station which complicates the solution of the problem and prohibits the use of the First-In-First-Out Principle (FIFO) principle; hence, its complexity is highly affected by the chosen model describing the transfers and their switching times. This problem is NP-Hard and the developed algorithms treat only small scale networks. For this reason, [104] focused on the modeling and the representation of data in order to allow the direct execution of Dijkstra's algorithm and other algorithms for the PPSPP while other researchers used bidirectional instead of unidirectional strategies to cover the graph or other heuristic methods.

[100]: Bertsekas (1993), 'A simple and fast label correcting algorithm for shortest paths'

[104]: Boardman et al. (1997), 'Computer assisted routing of intermodal shipments'

If one needs to compute a multi-modal fastest path which satisfies a predetermined set of users' preferences, the problem becomes even more complicated. The simplest case of the PPSPP under Constraints has been extensively studied and different methods for its solution have been presented based on the different types of parameters which demonstrate the users' preferences. The most computational friendly parameters have an assigned numerical value for each link, and are assumed totalizable along a path (e.g., the total distance of a path is the sum of the length of each link which consists the path). Hence, most of the developed algorithms consider a set of totalizable parameters which consists the users' preference.

The challenging task of computing the fastest path with respect not only to the travel time, but also to a number of constraints has been addressed by many different techniques. One solution for this general problem is to consider a generalised cost for every link of the network and perform Dijkstra's algorithm exactly as one would have done in the PPSPP. This generalised cost could contain the travel time of a link, the length of a link, the required fuel consumption, the greenhouse-gas emissions and other imposed parameters while each one of them is multiplied by a numerical value (weight) according to the users' preferences. The weight factor illustrates the level of user's preference for a distinct parameter. [105] proposed a utility measure for weighing the links with their cost, time or other constraints. However, even if the estimated function of the generalized cost simplifies the problem to a classical PPSPP, this method is not extensively used because it is indirect, it does not provide evidence that the selected path is the optimal one according to the user's needs and the weight factors are determined on an empirical basis.

Another, more straightforward method, allows the user to set an upper bound (threshold value) for each parameter which cannot be violated (e.g., the user may select to consume less than 8 litres of petrol and cover less than 65km during his trip to the joint leisure activity destination). In this case, the parameters are treated as secondary constraints and the objective function is constructed by the links' travel times. This problem is known as the Resource Constrained Shortest Path Problem (RCSPP) and is also an NP-Hard problem (see [106] and [107]); thus, it cannot be solved for relatively complex systems, like realistic transport networks. For this reason, the problem has been extensively studied and researchers have presented heuristic algorithms, which can be divided into the following two main categories:

[106]: Jaffe (1984), 'Algorithms for finding paths with multiple constraints'

Label Correcting Methods: These methods search exhaustively the whole network for an optimal solution by using the principles of Dynamic Programming (DP) or other heuristic methods. Because the network is exhaustively searched, pruning methods, such as the dominance method, have been developed to reject a number of options that cannot lead to an efficient solution with the intention of reducing the number of iterations and the complexity of the problem. Algorithms based on the principles of DP for the RCSPP were presented by [108], [109] and [110], while in [111] a performance assessment of three DP methods for the solution of the RCSPP is provided together with some comparative conclusions.

[111]: Righini et al. (2008), 'New dynamic programming algorithms for the resource constrained elementary shortest path problem'

Lagrangian Relaxation Methods: In these methods, the basic form of the RCSPP problem is changed. Generally, the secondary constraints of the RCSPP which form linear inequities are now introduced in the main objective function after being multiplied by the respective Lagrangian multipliers. Therefore, the objective function is changed and encompasses all the parameters along with the travel times of each link; therefore, algorithms for the PPSPP can be used to find a solution which minimizes the new objective function. A first algorithm which deals with the RCSPP under one constraint was presented by [112]. A Branch-and-Bound approach based on a Lagrangian heuristic for problems with many constraints was presented by [113] and was based on the well-known Subgradient method. [114] suggested the ellipsoid method to update the Lagrangian multipliers. Moreover, [115] investigated variants of the label-setting algorithm of [108] , and [116] proposed a suboptimal solution for the large-scale RCSPP problems

[108]: Desrochers et al. (1988), 'A reoptimization algorithm for the shortest path problem with time windows'

based on the extension to the discrete case of an exponential penalty function heuristic. [117] reduced the state space by aggregate resources and proposed an adjustment based on Lagrangian and surrogate relaxations in a column generation framework, and finally, [118] designed a three stage approach for column generation applications.

[118]: Zhu et al. (2007), 'Three-stage approaches for optimizing some variations of the resource constrained shortest-path sub-problem in a column generation context'

A final straightforward method is to consider the problem as a multi-objective shortest path problem where there is not a primary objective function (e.g., travel times) and secondary constraints which are expressed as linear inequalities, but a set of objective functions. The aim is to find all the efficient paths with the property that none of the objectives can be optimized further without worsening another one (Pareto Optimal Set of Paths). To demonstrate the difference of the RCSPP and the multi-objective SPP let consider the simple case of the bi-objective SPP with two positive costs $c_{ij} = \{c^1_{ij}, c^2_{ij}\}$ associated with each arc ij of the network, where i is the origin and j is the destination node and c^1_{ij} denotes the travel time of an arc and c^2_{ij} the length of the same arc. The length of an arc is a numeric totalizable parameter; hence if the user sets an upper bound for this parameter (e.g., b = 50km total covered distance), the RCSPP is formed as:

$$\begin{aligned} \text{minimize} \quad & \sum_{(i,j)\in p} c^1_{ij} \\ \text{subject to} \quad & \sum_{(i,j)\in p} c^2_{ij} \leq b \end{aligned} \tag{6.1}$$

while the bi-objective SPP is formed as:

$$\text{minimize } z(p) = \begin{cases} \sum_{(i,j)\in p} c^1_{ij} \\ \sum_{(i,j)\in p} c^2_{ij} \end{cases} \tag{6.2}$$

where p is a distinct path from the origin node, r, to the final destination, s, and $p \in P(r,s)$, where $P(r,s)$ is a set which contains all the feasible paths from r to s.

For the solution of the multi-objective SPP an extensive search in the network is also required and the problem is NP-Hard even in its bi-objective form [119] . This search is facilitated with labels and label-setting algorithms or label-correcting algorithms which have been developed for that purpose (a multiple objective extension of Dijkstra's algorithm can be also used). [120] provide a label-setting algorithm for the bi-objective SPP while the papers of [121], [122] and [123] address the multi-objective case. Label-collecting algorithms which are enumerate approaches are included in [124], [125] and [126] .

[119]: Serafini (1987), 'Some considerations about computational complexity for multi objective combinatorial problems'

[126]: Raith et al. (2009), 'A comparison of solution strategies for biobjective shortest path problems'

In the proposed mobile application the real-time multi-modal journey planning under users' preferences will be considered as a real-time multi-modal routing problem under a set of upper bounds for the secondary parameters which are determined by the users and should not be violated. As a result, the algorithms which deal with the users' preferences will address the RCSPP because: firstly it is computationally difficult to solve a multi-objective problem, secondly the user has always a predominant preference to select the fastest path and this is not facilitated in a multi-objective problem where all the parameters have the same value, and finally the proposed path is closer to the user's preferences if he denote precisely his will by setting an upper bound to his secondary constraints (e.g., willing to have less than 2 transfers and travel less than 50km).

6.3. Modeling the multi-modal journey planning under constraints

Let $G\{N, E\}$ be an urban transport network which is the outcome of the integration of different layers of a GIS system containing spatial information about different modes of transport. The set of vertices (intersections-stations) in the network is N and the set of arcs (links) that connect different stations is E. Each arc, $e \in E$, of the transport network has the following attributes:

- It is directed and connects an origin node $Or(e) = i \in N$ with a destination node $De(e) = z \in N$
- It can be traversed only by one transport mode and belongs to a distinct GIS layer: Layer(e)=k, where $k = 1, 2, ...$ is the number of the corresponding GIS layer (i.e., layer 1: passenger arcs, layer 2: private vehicle arcs, layer 3: rail arcs, etc.)
- It has an assigned travel time which varies according to the time when is traversed, $w(e, t - t_0) \in \mathbb{R}^+$. Hence, if t_0 is the current clock time, the travel time of the arc is relying on the time of traversing it (t). For a future time $t - t_0 \geq 0$, where the travel time of the arc is unknown, a traffic prediction tool is used
- It has a set of departure times denoting the time when the mode departs from the origin node of the arc, $e(\Delta T)$. For instance, a bus schedule which passes through an arc, e, is described as in Table 6.2

ΔT	1	2	3	4	5	6	7	...etc.
$E(\Delta T)$	12:00	12:10	12:18	12:35	13:01	13:18	13:36	...etc.

Table 6.2.: Description of a bus schedule passing through an arc e

This means that the arc $e \in E$ can only be traversed at a distinct time given by the time set $e(\Delta T)$ and the users have to wait till that time (this notation is especially used to describe the departure times of public transport modes).

Before proceeding to the network's description, it should be mentioned that two arcs may connect the same origin with the same destination, but the fact that they are describing the travel times and the departure times of different modes of transport make them distinct. Let now $i \in N$ be a node of the network with $\underbrace{\{a_1, a_2, a_3, a_4, ..., a_h\}}_{h=\text{total number of arcs}}$ outgoing arcs.

If a user arrives at node $i \in N$ at time $D(i)$ and he wants to traverse the arc $a_r \in \{a_1, a_2, a_3, a_4, ..., a_h\}$, then the Switching Time (ST) cost for the transfer from the current mode of transport to the transport mode, $Layer(a_r)$, is:

$$SW = min\{max\{a_r(\Delta) - D(i); 0\}\}\forall\Delta T \quad (6.3)$$

where $a_r(\Delta T)$ is the set of departure times for arc a_r. Following the above notation, if the user arrives at node $i \in N \mid Or(e) = i$ at time $D(i)$, then the required time to traverse an arc $e \in E$ -including the waiting time for departure- is:

$$min\{maxa_r(\Delta) - D(i); 0\} + w(e, min\{max\{a_r(\Delta T) - D(i); 0\}\})\forall\Delta T \quad (6.4)$$

A sequence of arcs $P = (e^1, e^2, \dots, e^x, e^{x+1}, ..., e^z)$, where $Or(e^{x+1}) = De(e^x)$ define a path from node $Or(e^1)$ to $De(e^z)$. Generally, an optimal path is acyclic and its nodes are visited only once because in transport networks $w(e, t) \geq 0$.

Let assume a travel from an origin node r to a final destination f with starting time t_0. It is also assumed that $P(r, f)$ is a subset of all feasible alternative paths from the origin to the final destination. If $p^* \in P(r, f)$ is the multi-modal fastest path from r to f, then p^* is derived by the following optimization problem,

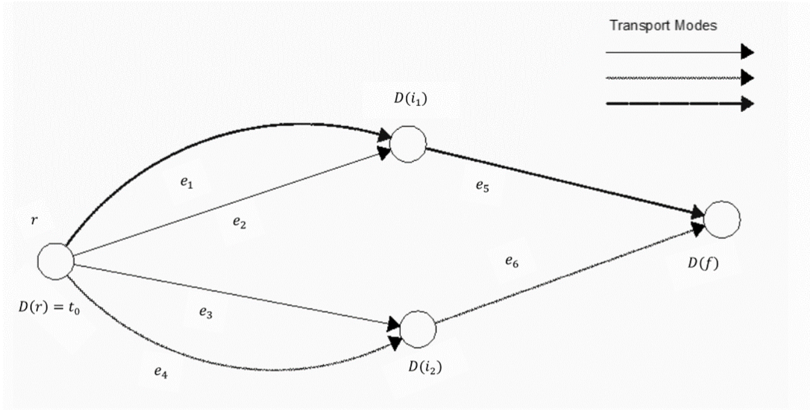
$$Minimise \sum_{e \in p}(\min\{\max\{a_r(\Delta T) - D(i); 0\}\} + w(e, min\{max\{a_r(\Delta T) - D(i); 0\}\})) \quad (6.5)$$

where $i = Or(e)$, $p \in P(r, f)$, $D(r) = t_0$ and $D(De(e)) = De(Or(e)) + min\{max\{e(\Delta T) - D(Or(e)); 0\}\} + w(e, min\{max\{e(\Delta T) - D(Or(e)); 0\}\})$.

If the fastest intermodal path is $p^* \in P(r, f) = (r, ..., e^x, e^{x+1}, ..., f)$, where $Or(e^{x+1}) = De(e^x)$, then if $Layer(e^x) \neq Layer(e^{x+1})$ a switch is made between different modes. Therefore, by performing that test for all the subsequent arcs from the beginning till the end of the path, one can derive the total number of transfers.

Description of the Multi-modal Journey Planning Problem under Constraints

A multi-modal journey with respect to a number of constraints is computed to satisfy the users' preferences or other potential local directives. In general, instead of providing data only for the travel times of the network's arcs to the central server, one can associate many other attributes to each arc of the network (e.g. information about CO_2 emissions in each arc of the network). In general, most of those attributes have numerical values, and are generally assumed totalizable along a path (e.g., an attribute of an arc describes the consumed amount of



$D(i_1) = \min\{ \min\{\max\{e_1(\Delta T)\text{-}D(r);0\}\} + w(e_1, \min\{max\{e_1(\Delta T) - D(r); 0\}\}) ; \min\{\max\{e_2(\Delta T)\text{-}D(r);0\}\} + w(e_2, \min\{\max\{e_2(\Delta T) - D(r); 0\}\}) \}.$
$D(i_2) = \min\{ \min\{\max\{e_3(\Delta T)\text{-}D(r);0\}\} + w(e_3, \min\{\max\{_3(\Delta T) - D(r); 0\}\}) ; \min\{\max\{e_4(\Delta T)\text{-}D(r);0\}\} + w(e_4, \min\{\max\{e_4(\Delta T) - D(r); 0\}\}) \}.$
$D(f) = \min\{ \max\{\min\{e_5(\Delta T)\text{-}D(i_1);0\}\} + w(e_5, \min\{\max\{e_5(\Delta T) - D(i_1); 0\}\}) ; \min\{\max\{e_6(\Delta T)\text{-}D(i_2);0\}\} + w(e_6, \min\{\max\{e_6(\Delta T) - D(i_2); 0\}\}) \}.$

FIGURE 6.2.: Representation of an Intermodal Network with three Alternative Mode Options

fuel while it is traversed and the fuel consumption of a path is the sum values of all arcs belonging to the path). Some classical totalizable attributes are the length of an arc, its travel time, its travel cost and its fuel requirements. Nevertheless, there are also variables with qualitative values which are not totalizable and are known as indexed. The fuel consumption and the covered distance are totalizable numerical variables with real positive values. If the fuel consumption is the attribute 1 of an arc and the total distance the attribute 2, then for an arc, $e \in E$, $R_e^1 \in R^+$ represents the fuel consumption and $R_e^2 \in R^+$ represents the covered distance while it is traversed. The number of transfers between different modes is totalizable, but it is not an attribute of an arc because it occurs after a transfer from one mode to another and this phenomenon cannot be explicitly described by the arc's attributes. Finally, the walking time is totalizable, and if it is considered as the 3rd attribute of an arc, then

$$R_e^3 = \begin{cases} R_e^3 \text{ if Layer(e)=1 (1 : passengers' layer)} \\ 0 \text{ if Layer(e) } \neq 1 \end{cases} \tag{6.6}$$

Initially, the more general form of the problem is considered, where a user has more than three totalizable preferences: $k = 1, 2, ..., m$. Let also $R_e^k \in R^+$ be the consumption of the resource $k = 1, 2, ..., m$ while traversing the arc $e \in E$ with mode $Layer(e)$. The consumption of each resource k while he traverses the path p by using a combination of modes is,

$$R_p^k = \sum_{e \in p} R_e^k, k = 1, ..., m \tag{6.7}$$

and if p is the path $p = (e^1, ..., e^x, e^{x+1}, ..., e^h)$, where $Or(e^(x + 1)) = De(e^x)$, then the total number of transfers, $TR \in Z^+$, is assumed initially equal to zero and its final value is derived by the search: for each $l = 1 : h - 1$, if $Layer(e^{l+1}) \neq Layer(e^l)$ then $TR = TR + 1$.

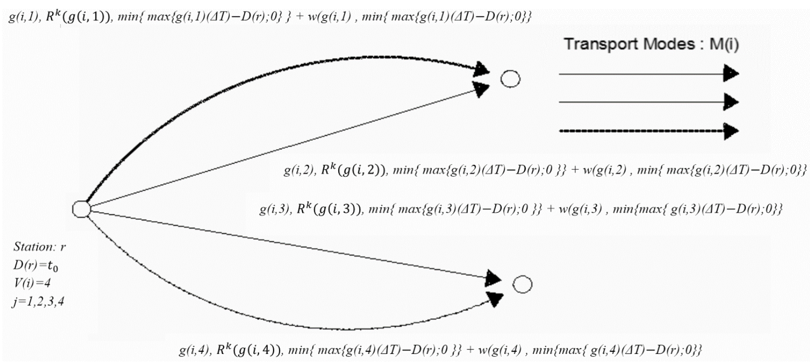


FIGURE 6.3.: Network's Representation as a directed graph

If one sets a series of upper bounds (thresholds), b^k, where $k = 1,2,3,...,m$ (e.g., maximum fuel consumption, maximum covered distance, maximum walking time, maximum value of other numerical totalizable attributes and finally maximum number of transfers permitted) for each one of the resources, then the resource consumption while traversing the path p should not violate any of these limitations. Assuming that a resource is consumed only if a link is traversed, then the Multi-Modal Shortest Path under Resource Constraints is a path $p^* \in P(r,f)$ which is derived from the optimization problem,

$$\begin{aligned} \text{minimize} \quad & \sum_{e \in p} (min\{max\{e(\Delta T) - D(i); 0\}\} + w(e, min\{max\{e(\Delta T) - D(i); 0\}\})) \\ \text{subject to} \quad & \sum_{e \in p_*} R_e^k \leq b^k, \forall k = 1, ..., m-1 \\ & TR \leq b^m \end{aligned} \tag{6.8}$$

A Heuristic Algorithm for the Mobile Application

Firstly, let define the outgoing arcs from each station $i \in N$ of the network. This is a trivial issue since the function $Or(e)$ determines the origin node of each arc $e \in E$ and only an initial search is required. Hence, for each node $i \in N$ its outgoing arcs $g(i,1), g(i,2), ..., g(i,V(i))$ are stored in a double matrix, $g(i,j)$, where the first term $i \in N$ denotes the origin node, the second term is a pointer $j \in 1,2,...,V(i)$ for the arc, and $V(i)$ is an integer value which denotes the maximum number of outgoing arcs from node i. Finally, for each $j \in 1,2,...,V(i), Or(g(i,j)) = i$. If an arc $g(i,j)$ is traversed, an amount of the totalizable resources $k = 1,2,...,m$ is consumed. Therefore, when an arc is traversed the resource consumption equals $R^k(g(i,j))$ for every resource $k = 1,2,...,m$. Moreover, when an arc is traversed after arriving at its origin node at time t_0, a travel time: $min\{max\{g(i,j)(\Delta T) - D(i); 0\}\} + w(g(i,j)\,, min\{max\{g(i,j)(\Delta T) - D(i); 0\}\}$ is required. In Fig.6.3, which represents a sketch intermodal network with three modes, the notation of each arc, the required resource consumption in order to traverse it and its anticipatory travel time are presented.

Consider that every path is described by a label containing a number of lists, where every list is a single dimensional matrix. The first list

$A : Z^+ \longrightarrow Z^+$, contains the stations $i \in N$ which are reached from the proposed algorithm during its computational procedure. Initially, the real location of the user, r, is pinpointed by his GPS receiver and stored in the list $A(1) = r$. The values of the list $A(1), A(2), A(3), ..., A(j), ...$ represent nodes in the network. The second list $D(A(j)) \in R^+$, contains the aggregated travel cost from the origin node r to node $A(j)$, where $A(j) \in N$. The other lists: $RC^k(A(j)) \in R^+, k = 1, 2, ..., m$, contain the total consumption of each totalizable resource k for the travel $r \longrightarrow A(j), A(j) \in N$. Finally, the list $TR(A(j))$ contains the total number of transfers which were made during the travel from r to $A(j)$. Generally, every label represents a different path from the origin station r to a significant place in the network, $A(j) \in N$. A typical label has the form $\{u, A(u), D(u), RC^k(u), TR(u)\}$ where $u \in \{1, 2, 3, ..., j, ...\}$ is a number which characterizes the label, $A(u)$ is a station, $D(u)$ represents the travel cost from the origin node r to $A(u)$, $RC^k(u)$, $k = 1, 2, ..., m$ represents the consumption of each resource k when this path is traversed and $TR(u)$ the total number of transfers. For instance, $\{1, r, t_0, 0, ..., 0\}$ is a label which represents the path $r \longrightarrow A(u) = r$ where $D(u) = t_0$, $RC^k(u) = 0, k = 1, 2, ..., m$ and $TR(u) = 0$.

Proposition 6.3.1 *A label $\{u1, A(u1), D(u1), RC^i(u1), TR(u1)\}$ dominates another label identified as $\{u2, A(u2), D(u2), RC^i(u2), TR(u2)\}$ only if the four below criteria are satisfied:*

- $A(u1) = A(u2)$ *"same node"*
- $D(u1) \leq D(u2)$
- $RC^i(u1) \leq RC^i(u2) \forall i = 1, ..., m.$
- $TR(u1) < TR(u2)$

Consequently, the second label can be permanently erased and this procedure is known as the "Domination Step".

The theory of domination is used in almost all algorithms which search for a solution on a problem via a label setting or a label correcting method, and it is important because it eliminates infeasible labels in the first place and reduces the storage requirements and the computational time. In the present case, one knows that route $r \longrightarrow A(u1)$ requires less travel time, consumes less resources and has less transfers than route $r \longrightarrow A(u2)$. Furthermore, the same modes depart from the stations $A(u1) = A(u2)$. As a result, for every route $p = \{A(u1) = A(u2), ..., x_k, x_{k+1}, ..., s\}$ from the new origin node $A(u1) = A(u2)$ to the destination node f, the route $r \longrightarrow A(u1) + p$ costs less and has lower resource consumption and number of transfers than route $r \longrightarrow A(u2) + p$.

Description of a solution: The proposed algorithm starts from the origin node and creates labels till the final destination is reached for the first time after a best-first search. Initially, the first label is $\{1, A(1), D(1), RC^k(1), TR(1)\} = \{1, r, t_0, 0, ..., 0\}$ and represents the loop less path $r \longrightarrow r$. Firstly, a list $Q = u$ is set which contains only one element: $u = 1$. Afterwards, the process starts from the origin node $A(u) = r$ and creates a new label for every outgoing arc $g(r, j)$ from node $A(u) = r$, where $j = 1, ..., V(A(u))$, and $V(A(u))$ is the number of outgoing links from node r. Hence, for every $j = 1, ..., V(A(u))$ is created a new label with the form number of label, destination node of arc j, travel time from origin to destination node of arc j, resource consumption from origin till the destination node of arc j, number of transfers from origin to destination node of arc j.

Let $g(A(u), 1)$ be the first egress arc from origin which creates the second label. The destination node of arc $g(A(u), 1)$ is $A(2) = De(g(A(u), 1)$. The travel time from the origin to the destination of this arc is $D(2) = D(u) + min\{max\{g(A(u), 1)(\Delta T) - D(u); 0\}\} + w(g(A(u), 1), min\{max\{g(A(u), 1)(\Delta T) - D(u); 0\}\})$. The resource consumption of each resource $y = 1, ..., m$ from the origin to the destination of arc is $RC^y(2) = RC^y(u) + R^y(g(A(u), 1))\forall y = 1, ..., m$. The number of transfers from the origin to the destination node of arc $g(A(u), 1)$ is $T(2) = 0$ because before the arc $g(A(u), 1)$ there is no other arc.

Hence, one derives the layer

$$\begin{aligned}\{2, De(g(A(u), 1), D(u) + min\{max\{g(A(u), 1)(\Delta T) - D(u); 0\}\} \\ + w(g(A(u), 1), min\{max\{g(A(u), 1)(\Delta T) - D(u); 0\}\}), \\ RC^y(u) + R^y(g(A(u), 1)), \forall y = 1, ..., m, T(A(u))\}\end{aligned}$$

The same procedure continues for every $j = 1, ..., V(A(u))$.

After this step, a new station $A(j) \neq A(u)$ which satisfies the conditions $D(j) \leq D(i), \forall i \cap Q = \varnothing$ and $RC^y(j) \leq b^y, \forall y = 1, ..., m$ and $TR(j) \leq b$ is marked ($b^y, \forall y = 1, ..., m$ are threshold values for each one of the resources and b is the number of permitted transfers).

The procedure is updated by setting $u \leftarrow j$ and $Q = Q + \{u\}$. The same procedure is repeated for every outgoing link from the new station $A(u)$ and new labels are created. From the second iteration till the end of the procedure a number of labels are eliminated if they are dominated. Hence, for each new label $\{v, A(v), D(v), RC^y(v), TR(v)\}$, we search if there is another label $\{v', A(v'), D(v'), RC^y(v'), TR(v')\}$, where $A(v) = A(v')$ and $D(v) \leq D(v')$, $RC^y(v) \leq RC^y(v'), \forall y = 1, ..., m$ and $TR(v) \leq TR(v')$ and if such a label already exists the new label is erased.

Finally, the procedure terminates when $u \leftarrow j_*$ and $A(u) = f$, where f is the final destination. The fastest path in an intermodal network under resource constraints is represented by the label $\{j_*, f, D(j_*), RC^y(j_*)\forall y = 1, ..., m, TR(j_*)\}$.

Lemma 6.3.2 *The multi-modal fastest path under constraints can be found when the final destination node f' is marked for the first time.*

Proof *If the node $A(j_*) = f$ is marked for the first time during that procedure, then $D(j_*) \leq D(i), \forall i \cap Q = \varnothing$ and $RC^y(j_*) \leq b^y \forall y = 1, ..., m$ and $TR(j_*) \leq b$. In most cases there are several alternatives that connect the origin with the destination and are represented by the labels $j_{**}, j_{***}, ...$ where $A(j_{**}) = A(j_{***}) = f$. However, these paths have not been marked yet. Therefore, $D(j_{**}), D(j_{***}), ... > D(j_*)$ or $RC^y(j_{**}), RC^y(j_{***}), ... > b^y, \forall y = 1, ..., m$ or $TR(j_{**}), TR(j_{***}), ... > b$ and, as a result, the fastest path under resource constraints is represented by the label j_*.*

Algorithm 4: Algorithm for multi-modal journey planning under resource constraints

A label $\{u, A(u), D(u), RC^y(u), \forall y = 1, 2, ..., m, TR(u)\} = \{1, r, t_0, 0, ..., 0\}$ represents the path $r \rightarrow r$, where $u = 1, A(u) = r, D(u) = t_0, RC^y(u) = 0, \forall y = 1, 2, ..., m$ and $TR(u) = 0$;
Set $J \leftarrow 0$ and $q \leftarrow 0$;
while *$A(u) \neq f$, where f is the final destination* **do**
  for *every arc $g(A(u), k)$ outgoing the node $A(u)$, where $k = 1, ..., V(A(u))$ and $V(A(u))$ is the total number of these arcs* **do**
    Set $J \leftarrow J + 1$;
    Set $E(J) \leftarrow g(A(u), k)$;
    $RC^y(J) \leftarrow RC^y(u) + R^y(g(A(u), k)), \forall y = 1, ..., m$;
    if *$q \neq 0$ and $Layer(g(A(u), k)) \neq Layer(q)$, then set $TR(J) \leftarrow TR(u) + 1$* **then**
      if *$RC^y(J) > b^y$, for one $y \in \{1, ..., m\}$, or $TR(J) > b$ then one of the constraints is violated. Therefore, we set $J \leftarrow J - 1$ and break the **until** loop, else set:* **then**
        $A(J) \leftarrow De(g(A(u), k))$;
        $D(J) \leftarrow D(u) + min\{max\{g(A(u), 1)(\Delta T) - D(u); 0\}\} + w(g(A(u), 1), min\{max\{g(A(u), 1)(\Delta T) - D(u); 0\}\})$;
        Set $i = \{1, ..., J - 1\}$;
        if there is a $z \in i \mid A(z) = A(J)$ and $D(z) \leq D(J)$ and $RC^y(z) \leq RC^y(J), \forall y = 1, ..., m$ and $TR(z) < TR(J)$, then this path is dominated, so set $J \leftarrow J - 1$ and return to the **until** loop, else:;
      end
    end
    Create a new label: $\{J, A(J), D(J), RC^y(J), \forall y = 1, ..., m\}$;
  end
  Set $H = \{1, ..., J\} - Q$;
  Find $J \in H$ for which $D(J) \leq D(i) \forall i \in H$ and set $q \leftarrow E(J)$ and $u \leftarrow J$;
end
Set $Q \leftarrow Q + \{u\}$;

The complexity of Algorithm 4 is $O(\phi J)$, where ϕ is the number of total iterations and J the number of created labels. The values of these two parameters depend on the number of stations in the network, the number of arcs, the different modes and especially from the threshold values of the secondary constraints; hence one cannot determine them in the first place. For this reason, in order to minimize the running time of the algorithm, the number of total iterations should remain in low levels.

Pruning Labels

The total number of created labels could be reduced if one introduces more criteria for the elimination of the created labels in the first place. In the proposed pruning method, for each totalizable resource parameter $k = 1, 2, ..., m$ a simple PPSPP algorithm from every station of the network to the destination node is performed and the values $FR^k(i), \forall k = 1, 2, ..., m$ which represent the minimum possible resource consumption for the travel from each node $i \in N$ to the final destination f are computed. If Dijkstra's algorithm is used with a Fibonacci heap data structure, as presented by [127] , the pre-processing method has a running time complexity of $O(m(N((E + NlogN)))$, where m is the number of totalizable resource parameters N the number of network's nodes and E the number of arcs.

[127]: Fredman et al. (1987), 'Fibonacci heaps and their uses in improved network optimization algorithms'

Lemma 6.3.3 *A label* $\{u1, A(u1), D(u1), RC^k(u1), TR(u1)\}$ *is eliminated if the following criterion is satisfied:* $RC^k(u1) + FR^k(A(u1)) > b^k$ *for even one* $k = 1, ..., m.$ *, where* b^k *is the upper bound of the* k^{th} *totalizable resource parameter.*

Proof. *The proof is straightforward because if the resource consumption from the beginning till node* $A(u1)$*, denoted as* $RC^k(u1)$*, plus the minimum resource consumption required to complete the trip from node* $A(u1)$ *till the final destination, denoted as* $FR^k(A(u1))$*, is higher than the upper bound* b^k *for even one of the constraints, then this path will violate the upper bound (threshold) before it reaches the final destination and it can be eliminated.*

Reducing the Running Time by Minimizing the Searching Space

The proposed method computes a real-time sub-optimal path in a multi-modal network under a number of users' preferences. The path is called sub-optimal because it may not be the fastest path from an origin to a destination. However, the structure of the method ensures that the sub-optimal path will be either the fastest path or a path with a small time difference from the fastest one. In this method a trade-off between the complexity of the algorithm and the accuracy of the solution is made, but it is ensured that the accuracy loss is minimum whereas the computational benefit is high.

Firstly, the mean travel times for every arc in the network, is computed. If a day period is considered, then the mean travel time for an arc, e, is:

$$W(e) = \frac{\sum_{t=1}^{Y} w(e,t)}{Y} \tag{6.9}$$

where t is the number of travel-time updates in arc e during a day. The mean travel time of each arc could be also derived by historical data which contains more measurements.

Then a pre-processing algorithm is used in order to compute the fastest path from each node i in the network till the final destination f. Dijkstra's algorithm could be used for that purpose and with a total running time of $O(N(E + NlogN))$. The fastest path from each every station, $i \in N$, till the final destination is described by a value $H(i)$ which denotes the minimum possible cost for the travel from station i till the destination f.

The values $H(i)$ are not computed via real-time data and the mean arc travel times $W(e)$ were used during their computation instead of real-time travel times. Moreover, the switching times between different modes were also ignored during that computation. The aim of this pre-processing algorithm is to provide an estimation of the minimum required travel time to travel from each node of the network till the final destination. This estimation will be used at a later stage to reduce the searching space of the main algorithm and minimize the number of created labels.

Following the pre-processing algorithm, algorithm 4 is implemented with the add-in label-elimination criterion which was described in the previous section. However, pruning reduces further the searching space because in every iteration it is not the closest station that is selected from the origin $D(u) = minD(i), \forall i \cap Q = \varnothing$ for which $RC^y(u) \leq b^y, \forall y = 1, ..., m$ and $TR(u) \leq b$, but the station which satisfies the expression $\{D(u) + H(A(u)) = min(D(i) + H(A(i)), \forall i \cap Q = \varnothing$ for $RC^y(u) \leq b^y, \forall y = 1, ..., m$ and $TR(u) \leq b\}$, where $D(u)$ is the minimum real-time travel time which is required to traverse the path between the origin node r and the current station $A(u)$ including the switching times between different modes. After a number of iterations, the final destination will be marked for the first time $A(j_*) = f$, when $D(j_*) + H(A(j)_*)) \leq D(i) + H(A(i)), \forall i \cap Q = \varnothing, RC^y(j_*) \leq b^y \forall y = 1, ..., m$ and $TR(j_*) \leq b$.

In this stage it is known that $H(A(j)_*) = 0$, because $A(j)_*$ is the final destination f and the estimated travel time from node f to node f is equal to zero by definition. Hence, that is a first multi-modal path from origin to destination which satisfies the users' preferences and is denoted by the label $\{j_*, f, D(j_*), RC^y(j_*) \forall y = 1, ..., m, TR(j_*)\}$. It is not certain that this path is also the fastest one, but the fact that exists a path from r to f which requires a travel time $D(j_*)$ enables the elimination of all the labels which have a higher travel time than $D(j_*)$, because they have not reached yet the final destination and have a higher travel time than $D(j_*)$.

In the following step, j_* is added in the set Q, and the algorithm continues like before by selecting another label l, where $\{D(l) + H(A(l)) = min(D(i) + H(A(i)), \forall i \cap Q = \varnothing$ for $RC^y(l) \leq b^y, \forall y = 1, ..., m$ and $TR(l) \leq b$. After a number of iterations, the final destination will be marked for a second time $A(j_{**}) = f$, when $D(j_{**}) + H(A(j_{**})) \leq D(i) + H(A(i)), \forall i \cap Q = \varnothing, RC^y(j_{**}) \leq b^y, \forall y = 1, ..., m$ and $TR(j_{**}) \leq b$. If the real-time multi-modal travel time $D(j_{**})$ is less than $D(j_*)$, then it is assumed that the fastest path is now denoted by the label $\{j_{**}, f, D(j_{**}), RC^y(j_{**}) \forall y = 1, ..., m, TR(j_{**})\}$. In the following step, j_{**} is also added in the set Q, and the algorithm continues like before by selecting another label l, where $\{D(l) + H(A(l)) = min(D(i) + H(A(i)), \forall i \cap Q = \varnothing$ for $RC^y(l) \leq b^y, \forall y = 1, ..., m$ and $TR(l) \leq b$.

The whole procedure continues iteratively until the final destination is marked for the tenth time. After that the fastest path from a list of ten candidates is assumed to be the one which required the less real-time multi-modal travel time to reach the destination. If the procedure continues more than ten times, one can be more confident that the proposed path is also the fastest path and the number of iterations is related to the level of accuracy. However, the first iterations improve markedly the level of accuracy and after a while the proposed paths from additional iterations are too time-consuming. The accuracy of the estimations $H(i)$ plays also an important role to the final solution because if they are similar to the real-time data outcomes, the fastest path will be found only after a small number of iterations.

6.4. Numerical Experiments

This section summarizes the performance of the proposed algorithmic framework for the selection of a real-time multi-modal fastest path under a number of constraints for moving from an origin to a joint leisure activity location. The algorithms were tested on a reference 2556 MHz-processor machine with 1024 Megabytes RAM. The tests were implemented on randomized networks that contained 100, 200, 300, 400, 500, ..., 3000 nodes-stations. In order to minimize the effects due to network's topology, both sparse and dense networks were examined. The execution environment of each test is described by the number of nodes, the number of arcs, the level of density, the number of transportation modes and the number of constraints. Each arc connects an origin with a destination node via a distinct transport mode and more arcs than one may connect the same origin-destination pair.

In Table 6.3, detailed results are given with respect to one constraint. Details are given for the number of nodes-stations (column 1.), the number of arcs (column 2.), the density of the network (column 3.), the number of different available modes in the network (column 4.), the constraints upper bound (column 6.), the total travel time and the resource consumption of the fastest path (column 7.), the running time of algorithm 4 (column 8.), the number of created labels with algorithm 4. (column 9.), the running time of algorithm 4′ (which is the extension of algorithm 4 proposed at the previous section) (column 10.), the number of created labels with algorithm 4′ (column 11.) and the accuracy of algorithm 4′ (column 12.). The accuracy, A, of algorithm 4′ is calculated as:

$$A = (1 - \frac{D - D^{'}}{D})100\% \tag{6.10}$$

where D is the travel time of the fastest path from origin to destination and $D^{'}$ is the travel time of the path which is proposed by algorithm 4′.

In Table 6.4, the results of several alternative networks are presented with respect to four constraints (fuel consumption, covered distance, walking time and number of transfers). On the contrary, the performance of algorithm 4 is not presented in Table 6.4, for the simple reason that its running time exceeds the predetermined running time threshold of 200 sec.

Table 6.3.: Running times and labels under one constraint

Randomised Network			Number of Modes	Number of Constraints	Constraint Limit (units)	Travel time (min) and resource consumption (units)	Complexity				
Intersections/ Stations	Purpose-Built Arcs	Ratio Nodes/Edges					Algorithm 4 Running Time (sec)	Number of Labels	Algorithm 4' Running Time (sec)	Number of Labels	Accu-acy
100	1000	10.00%	2	1	670	54.78 / 488.23	0.717	1294	0.0321	472	97%
200	2158	9.27%	3	1	1540	139.72 / 1486	11.844	5511	0.0349	2412	94%
600	49811	1.20%	3	1	430	12.3 / 234.5	7.22	3014	0.1013	1158	100%
1400	13250	10.57%	4	1	1021.76	315 / 1011	>200	Out of bound	0.8512	23225	94%
1600	15420	10.38%	4	1	972.72	342 / 963.5	>200	Out of bound	0.4964	15942	92%
1800	17690	10.18%	5	1	861.85	369 / 850.4	>200	Out of bound	1.325	17526	96%
2600	24980	10.41%	5	1	727.61	358 / 716.1	>200	Out of bound	1.748	16954	97%
500	25200	1.98%	3	1	415.17	36 / 405.1	86.12	58241	0.1491	4223	100%
300	20434	1.47%	3	1	588.41	19 / 567.3	0.1748	2981	0.071	392	100%
400	23962	1.67%	4	1	427.22	27 / 421.11	0.336	3769	0.093	612	100%
2800	27312	10.25%	6	1	911.61	334 / 906.3	>200	Out of bound	0.896	9125	93%
3000	29762	10.08%	6	1	1052.8	315 / 1044.16	>200	Out of bound	0.847	7032	95%

Table 6.4.: Running times and labels under four constraints

Randomised Network			Number of Modes	Number of Constraints	Limits of Constraints (units)	Travel Time (min)	Complexity		Accuracy
Intersections/ Station	Purpose-Built Arcs	Ratio Nodes/Edges					Algorithm 4' Run. Time (sec)	Number of Labels	
100	1000	10.00%	2	4	15.6/135/19/1	79	0.0321	2113	100%
200	2009	9.95%	3	4	13.6/115/17/2	58	0.0812	8051	100%
1000	11000	9.25%	3	4	15.4/122.9/15/5	61	0.1932	9736	100%
1400	14000	10.57%	4	4	22.5/186/13/4	228	34.11	379125	91%
1600	16000	10.38%	4	4	31/285/25/2	181	7.048	135209	96%
1800	18000	10.18%	5	4	66/470/34/3	439	2.311	51611	92%
2600	26000	10.41%	5	4	200/1550/18/2	874	3.72	85236	95%
500	26000	1.98%	3	4	10.7/91/11/3	89	0.142	6948	100%
300	20400	1.47%	3	4	2/12.4/21/2	28	0.314	307	100%
400	24000	1.67%	4	4	6/39.6/23/3	49	0.307	811	100%
2800	28000	10.25%	6	4	89/540/14/4	412	21.49	192597	92%
3000	30000	10.08%	6	4	161/1040/28/3	617	4.07	51321	96%

In addition, as was already discussed, the running time complexity of algorithm 4 is $O(\phi J)$, where ϕ is the number of total iterations and J the number of created labels. The complexity of the algorithm cannot be directly associated with the number of nodes or arcs because the total number of created labels is affected by many other secondary parameters (e.g. the upper bounds of constraints, the domination criteria, the nature of the examined network and more). For this reason, in Figure 6.4, a diagram is shown which demonstrates the relationship between the algorithm 4 running time in sec. and the number of created labels which implies a linear pattern between them. Figure 6.5, shows a plot of the number of nodes multiplied by the number of arcs versus the total number of created labels which suggests that a strong relationship between them cannot be established because secondary parameters play also an important role.

The proposed mobile application is based on those algorithms in order to compute the optimal path of joint leisure activity travelers. The numerical experiments showed that the algorithm 4 has an acceptable performance for medium-sized networks (e.g., 600 stations, 49,811 links, 3 different modes in the network, 1 constraint or 500 stations, 25,200 links, 3 modes, 1 constraint), but the complexity of the problem and the technological constraints inhibit, as we speak, the use of this algorithm for large networks with few constraints. As a result, other heuristic approaches which have an acceptable performance in larger networks with multiple constraints (such as algorithm 4′ that includes also pruning) can be used as an alternative, even if they yield sub-optimal solutions. Algorithm 4′, performs well in large networks under multiple-user constraints (e.g. 2800 nodes, 28,000 arcs under four constraints). Moreover, the level of accuracy falls between 90-100%, which means that in the worst case, the sub-optimal path may differ from the fastest one up to 3 minutes for small trips (from 20 to

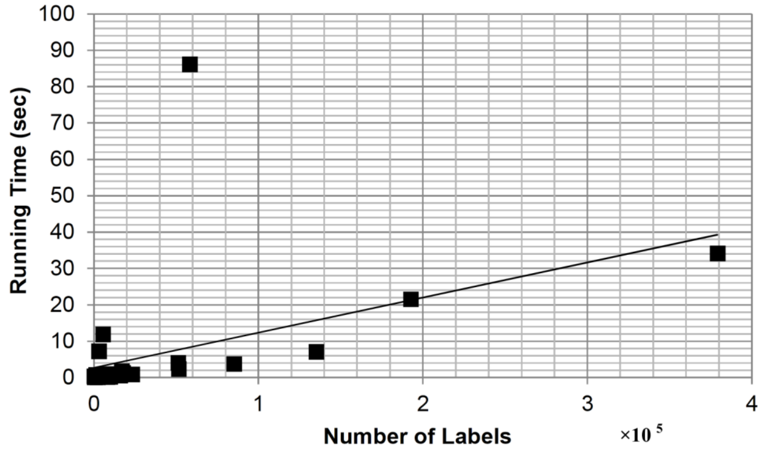


FIGURE 6.4.: Running time versus number of labels

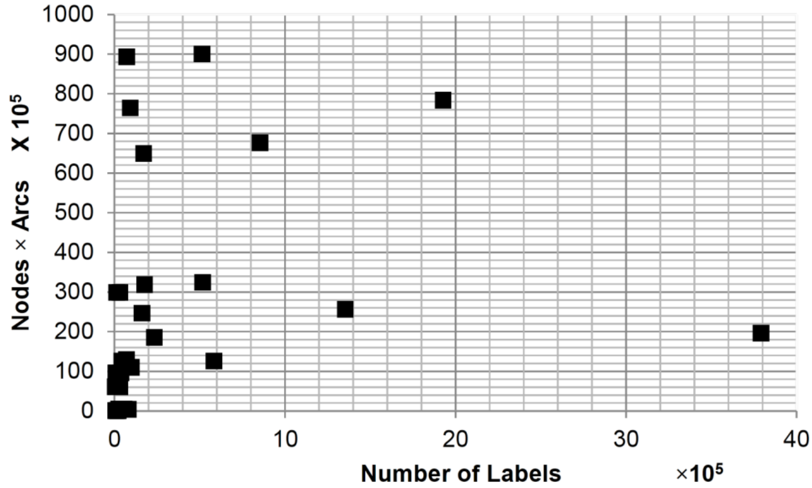


FIGURE 6.5.: Nodes × Arcs versus number of labels

30 minutes) and up to 9 minutes for a 90-minute trip which could be considered near the upper limit for an urban network.

Car Sharing for attending Joint Leisure Activities

7.

At the previous chapter, we studied how the best journey from one origin point to the destination of a joint leisure activity can be suggested to each user considering multiple modes of transportation and users' preferences. At this approach, there is no collaboration among the users that participate at the same joint leisure activity since everyone tries to optimize his/her travel based on his/her own needs considering all available modal options.

In this chapter, a collaborative approach is explored where some of the users who are willing to participate at a joint leisure activity own a private vehicle and can pick-up some of the other users that will participate also at the same activity. Under this collaborative car-sharing scheme, the generation of unnecessary trips can be reduced.

Because users travel towards the same joint leisure activity destination, trips can be spared only if multiple users share the same transport mode (i.e., private car). For this reason, this chapter focuses solely on the use of car-sharing from several users without considering other alternative means of transportation (i.e., multi-modality).

7.1. Modeling the Car-sharing problem for Joint Leisure Activity Participation

Let assume that a number of individuals $i = \{1, 2, ..., |i|\}$ are willing to participate at the same activity which is located at one point B and a fraction of those individuals possesses a private car. To model the car possession, a dummy variable y_i is assigned to each individual where:

$$y_i = \begin{cases} 0 : \text{if individual } i \text{ does not posses a car} \\ 1 : \text{otherwise} \end{cases}$$

The full list of sets, subscripts, parameters and variables used in the modeling of the car sharing problem for joint leisure activity participation is presented below.

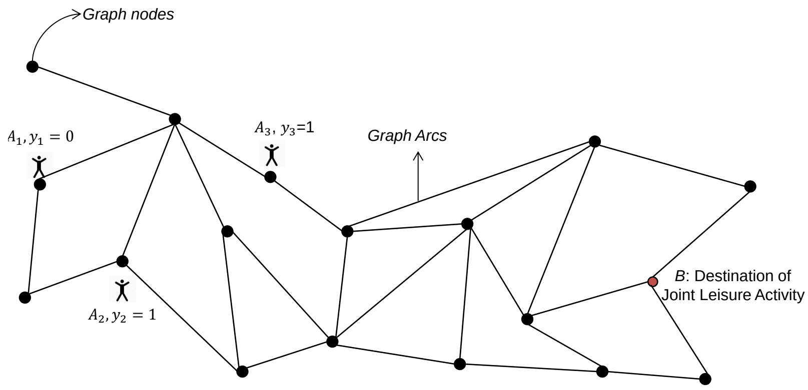


FIGURE 7.1.: Example of graph representation $G\{N, E\}$ with three (3) individuals willing to attend a joint leisure activity

$i = \{1, 2, ..., \lvert i \rvert\}$	the number of individuals participating at a joint leisure activity
y_i	a dummy binary variable denoting the possession of a private car for each individual i
$G\{n, e\}$	a graph representation of the urban road network
$n = \{1, 2, ..., \lvert n \rvert\}$	the number of nodes at the urban network (i.e., road intersections)
$e = \{1, 2, ..., \lvert e \rvert\}$	the number of arcs at the urban network connecting nodes
O	a vector with $\lvert e \rvert$ elements denoting the origin node of each arc: $O_e \in n$
D	a vector with $\lvert e \rvert$ elements denoting the destination node of each arc: $D_e \in n$
t	a vector with $\lvert e \rvert$ non-negative elements denoting the travel time (or travel cost) of traversing each arc, t_e
A	a vector with $\lvert i \rvert$ elements denoting the origin location of each individual i from which he/she has to start his/her journey towards the joint leisure activity location B
U	a scalar positive value denoting the gained utility in case joint leisure activity participants share the same car for their journey

A representation of this problem is provided in Fig. 7.1.

In the simple case that one user has a private vehicle and is responsible of picking up every other user who will participate at the joint leisure activity with him, computing the optimal route is equivalent of solving the well-known *traveling-salesman* problem. Nevertheless, some of the users might not be picked up in the end if the traveling cost of picking them up exceeds the gained utility benefit, U, of reducing the number of trips.

This leads to the introduction of an objective function of the form:

$$min \sum_{e} l_e t_e - k \times U \tag{7.1}$$

where l_e is an integer number denoting how many times arc e was traversed from users traveling to the joint leisure activity destination (it can be also $l_e = 0$ if nobody traversed that arc) and k a scalar vector that denotes how many trips were shared by the same car.

Finding the optimal values of the solution variables of the problem ($l_e \forall e$ and k) is not a trivial task since a number of secondary constraints should also be satisfied at the same time (i.e., all paths should start from the current location of one user i and end at the location of another user in case of pick-up or at the location of the joint leisure activity).

7.2. Dynamic Programming search of the best car-sharing option

The well-known dynamic programming algorithm of *Floyd-Warshall* is used for solving the all-pairs Shortest Path problem with computational cost $O(n^3)$, where n is the number of nodes. Then, $T_{k,l}$ denotes the computed minimum travel cost between each pair of nodes k, l.

After having computed the all-pairs shortest paths, the minimum cost of each individual i traveling alone to the destination of the joint leisure activity in case he/she uses a car is $T_{A_i,B}$ and the overall cost of all journeys related to that joint leisure activity is:

$$C = -k \times U + \sum_{i=\{1,2,\ldots,|i|\}} T_{A_i,B}$$

(7.2)

where k is the number of shared trips, which in this case are equal to zero because each individual traveled alone.

Starting from there, the objective is to find if the overall cost C can be reduced further if a number of individuals share the same car. For checking this, one should start from one location A_i and check all possible pick-up combinations requiring a total search of $|i| \times (|i| - 1)^2$ options in the worst case that all individuals have private cars and all starting points should be checked.

For instance, let assume that one user $i = 1$ has a private vehicle and one is willing to check the cost of the alternative of picking up another user $i = 3$. For this, the new total cost is:

$$C = -k \times U + \left(\sum_{i-\{1,3\}} T_{A_i,B} \right) + 2T_{A_3,B} + T_{A_1,A_3} \tag{7.3}$$

where T_{A_1,A_3} is the minimum picking up cost of individual $i = 3$ from individual $i = 1$ and $T_{A_3,B}$ the remaining travel cost from location A_3 to the location of the final activity. In addition, we have one joint trip; thus, $k = 1$, and if the gained utility $k \times U$ is greater than the additional inflicted travel cost: $(T_{A_1,A_3} + T_{A_3,B}) - T_{A_1,B}$, the new total cost is lower than the previous case.

For checking all possible pick-up combination options, $|i| \times (|i| - 1)^2$, the number of computations is at the polynomial level; thus, ensuring the problem scalability. The proposed algorithm for suggesting the optimal car-sharing option to joint leisure activity participants is presented below together with its associated computational cost for different number of participants (Fig.7.2).

Algorithm 5: Compute the best car-sharing option for a group of joint leisure activity participants including the order of pick ups with polynomial computational cost: $O(|n|^3 + |i| \times (|i| - 1)^2)$

Set $T_{z_1,z_2} = +\infty$ for all $z_1, z_2 \in n$;
for *each* $e \in \{1, 2, ..., |e|\}$ **do**
    $T_{O_e,D_e} = t_e$
end
Use *Floyd-Warshall* algorithm to compute all-pairs shortest paths;
for *all* $z_1 \in \{1, 2, ..., |n|\}$ **do**
    for *all* $z_2 \in \{1, 2, ..., |n|\}$ **do**
        for *all* $z_3 \in \{1, 2, ..., |n|\}$ **do**
            $T_{z_2,z_3} = min(T_{z_2,z_1}; T_{z_2,z_1} + T_{z_1,z_3})$;
        end
    end
end
Compute $C = -k \times U + \sum_{i=\{1,2,...,|i|\}} T_{A_i,B}$;
for *all users i willing to participate at the joint leisure activity* **do**
    if $y_i = 1$ *(user i has a private vehicle)* **then**
        for *each alternative pick-up combination of user i* **do**
            Compute the new total cost C' and if $C' < C$, then set $C \leftarrow C'$;
        end
    end
end
Return the car-sharing option (if any) with the lowest total cost C;

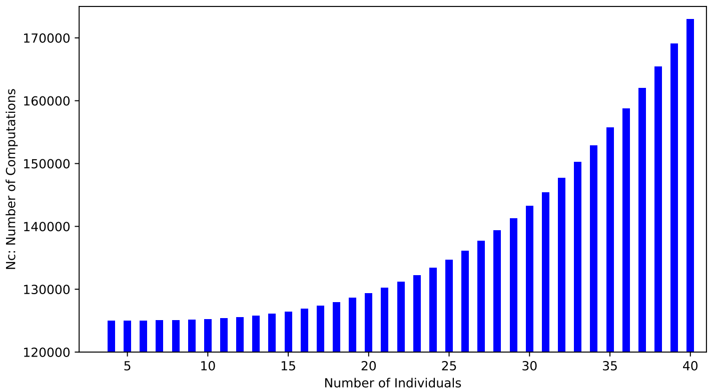


FIGURE 7.2.: Required number of computations for a network with 50 arcs

8. Conclusion

8.1. Summary of Thesis Contribution

This thesis followed a holistic approach on the problem of trip optimization for participating at joint activities. The studied problems were split in five interdependent categories which were rigorously analyzed at the previous chapters:

1. Understand the SoA work on utilizing user-generated data for increasing the efficiency level of joint leisure activities and propose actions towards this direction.
   thesis contribution: the utilization of (i) Cellular Data (ii) Social Media Data (iii) Smart Card Data and (iv) Geo-location Data from PDAs was rigorously studied together with the current SoA applications. During the exploration of non-recurrent activities, it was observed that although the full information for forming a decision-making objective function is obtainable, research works have not been focusing on that direction.
2. Capture users' willingness to travel certain distances for participating in different types of activities
   thesis contribution: a mobility pattern recognition model introduced for retrieving automatically users' mobility and activity patterns based on spatio-temporal analysis of historic user-generated geo-location data (Twitter data from London over a 14-month period). The frequency level of visiting a particular location was utilized to link locations with activity types and rules for linking one revisited location with home were introduced demonstrating an observed accuracy of more than 90%.
   A utility-maximization model was also introduced for capturing users' willingness to travel certain distances for participating in different types of activities for different day times and types. The model learned automatically the users' habits from historic data and offered valuable insights that can be used as source of information for suggesting common activities to multiple users. For performing such action, users' were clustered to capture the similarities on their mobility and activity patterns along with their willingness to travel similar distances to participate in certain types of activities. A step-by-step single-point estimate approach was also introduced for simulating individuals' daily schedule and identifying automatically locations for performing joint activities.
3. Optimize the selection of locations and starting times of joint leisure activities
   thesis contribution: the problem formulation of the optimization problem considering the willingness of users to travel certain distances to participate at different types of activities was introduced for the first time. Then, evolutionary optimization

was utilized for confronting the scalability problem of the Joint Leisure Activity optimization. The problem-specific stochastic annealing heuristic reduced the exponential computational complexity and due to the improved scalability, this approach enables the development of Web and Smartphone applications for suggesting automatically the time and the location for performing a joint leisure activity. In a practical use-case, after applying the stochastic annealing search for optimizing 9 Joint Leisure Activity Instances, the perceived utility of users was increased up to 3 times compared to the basic scenario because of the rescheduling of the arrival times of individuals and the activity locations.
Due to absence of alternative heuristics, the stochastic annealing method was compared against problem-tailored algorithms from the areas of genetic algorithms and hill climbing. Stochastic annealing was capable of finding approximations to the optimal solution that increased up to 2 times the perceived utility of users in a cluster compared to the hill climbing and the GA.

4. Re-schedule public transportation in near real time in order to adjust to the joint leisure activity demand without deteriorating QoS for other passengers
thesis contribution: the problem of public transport re-scheduling for different public transport services that can cover joint leisure activity passenger demand subject to the no-deterioration of service quality and the adherence to a set of operational regulations was modeled. For this, a sequential heuristic search algorithm for changing the public transport schedules in near real time (in a matter of minutes) for adjusting to the arrival time needs of joint leisure activity passengers without deteriorating the KPIs of public transport operations was introduced.
In a test-case, GTFS data from Sweden focusing on bus lines 1 and 4 in Stockholm and Twitter data for deriving individual trips to joint leisure activity locations in Stockholm were utilized . Due to the schedule changes after optimization, the operational performance of bus services demonstrated an EWT improvement at a service-wide level for all services while only some stations from line 1, direction 2 and line 4, direction 4 had a slight EWT deterioration(up to 0.6min.); without affecting significantly though the level of service of bus operations. At the same time, after the re-scheduling, the joint leisure activity trips from all services enjoyed new arrival times to the joint activity stations which were more than 50% closer to the starting times of those activities. Finally, the computational cost of the proposed heuristic algorithm for public transportation re-scheduling demonstrated that a convergence to an approximate global optimum requires from 2-6 minutes.
5. Optimize the journey selection of users' who are willing to travel from one point of the network to another for participating at one activity and, possibly, utilize multiple modes while also satisfying their preferences
thesis contribution: a mobile phone application was described for attempting to establish a unified approach for mobile navigation with continuous data feeds from different sources. This

application is based on complex algorithms in order to compute the optimal path due to the multi-modal character of the path selection problem and the resource constraints that emerge from users' preferences. For this, a fastest-path-tree algorithm that converged to the global optimum under a set resource constraints that considers multiple modes with the use of GIS layers was developed and demonstrated an acceptable performance for medium-sized networks (e.g., 600 stations, 49,811 links, 3 different modes in the network, 1 constraint or 500 stations, 25,200 links, 3 modes, 1 constraint). However, the complexity of the problem and the technological constraints inhibit, as we speak, the use of this algorithm for large networks with few constraints. As a result, a pruning extension was added to this heuristic approach which demonstrated an acceptable performance in larger networks with multiple constraints with the drawback that it cannot guarantee the finding of the global optimum. In practice, it performed well in large networks under multiple-user constraints (e.g. 2800 nodes, 28,000 arcs under four constraints). Moreover, the level of accuracy was above 90%, which means that in the worst case, the suboptimal path may differ from the fastest one up to 3 minutes for small trips (from 20 to 30 minutes) and up to 9 minutes for a 90-minute trip which could be considered near the upper limit for an urban network.

The thesis offered an important gain to society as additional light was shed on capturing users' mobility patterns/preferences with the use of user-generated data and optimizing non-recurrent trips which can be more than 60% of the total number of trips in cities. Individual users may become able to enumerate all potential activity alternatives in the city and choose in space and time the ones that maximize their utility without incurring high travel costs. They may also be able to share transport modes when they participate to joint activities and select the best journey option over a broad set of alternative paths and modes according to the (i) total journey time and (ii) their specific preferences. In addition, public transport operators may improve the competitiveness of their services by following a demand-responsive approach where they reschedule their operations for covering the joint leisure activity demand without penalizing their overall QoS. Apart from the transport operators, the transport network in urban environments may enjoy a significant improvement in traffic conditions if the non-recurrent trips which vary broadly from day to day are tackled in a more efficient way from travelers.

8.2. Discussion on Validation

The developed methodologies of this thesis were implemented with the use of datasets from different study areas which are summarized at Table 8.1. At this section the validation results are summarized in order to discuss the practice readiness of the developed methodologies at Chapters 3, 4, 5, 6.

Table 8.1.: Utilized Data Sources for Methodological Validation

Capturing users' willingness to travel certain distances for participating at different types of activities	user-generated tweets with geo-tagged locations from 65 twitter users in London over a 14-month period
Optimizing the Location and Time of Joint Leisure Activities	user-generated tweets with geo-tagged locations from 75 twitter users in London over a 14-month period
Adapting the departure times of Public Transportation services to the Joint Leisure Activity Demand	GTFS data from two bi-directional central bus lines in Stockholm, Sweden together with the list of operational constraints and **Social Media Data**, user-generated Data of 62 persons in the study area
Multi-Modal Journey Planning of Joint Leisure Activity Trips subject to personalized preferences	Topology of a study area. Travel times of all links in the study area for all alternative transport mode choices (i.e., bus, private vehicles etc.). Link traversing cost and fuel consumption of each link given the utilized transport mode. Preferences of users' undertaking trips together with their origin-destination points

In Chapter 3, a daily pattern recognition model was introduced for deriving the willingness of users to travel certain distances for participating at different types of activities. The daily pattern recognition model was tested with the use of a 14-month period Twitter data from 65 users and demonstrated a 90% accuracy on allocating re-visited locations with meaningful activities. The daily pattern recognition model is individual-based and is automatically updated when new user-generated data is provided. On top of the daily pattern recognition model, a utility-maximization model of each examined user was developed and it returned the level of willingness to travel certain distances for participating at different types of activities. However, there are two underlying assumptions on this generalization which were not possible to be validated given the data at hand. The first assumption was that the utility of the user depends on the perceived utility at performing an activity type and the traveled distance; however, the real location of that activity might play an important role (i.e., the user might not feeling comfortable with certain areas of the city). This cannot be validated in practice though if revealed preference surveys are not conducted. The second assumption was that all activities that belong to the same activity type (i.e., all restaurants) have the same effect on the perceived utility of a user. This assumption again cannot be validated because the user should provide information about how he/she ranks all the places of interest in the city. As with most individual-based utility-maximization models though, there should be some assumptions about the user's behavior since the full decision-making mechanism of an individual cannot be replicated with 100% accuracy. **Therefore, this method might require certain modifications when utilized in practice based on the data that is available at each practical implementation scenario**.

Those utility-maximization models for deriving the willingness of the users to travel certain distances to participate at different types of activities given the time of the day were utilized in Chapter 4 for selecting the optimal location and time of a joint leisure activity. Given the vast number of alternative activity locations in urban areas, heuristic search methods were developed such as the stochastic annealing search which converged at an optimal location and starting time for the joint activity. The stochastic annealing search produced stable results when tested with data from 10 user groups derived from 75 Social Media accounts. However, it was not possible to define how close to the global optimum were those results since the optimization problem was computational intractable given its exponential computational complexity. Given that, the stochastic annealing search method was tested against other SoA heuristics such as Hill Climbing and Genetic Algorithms and outperformed them by improving up to 2 times the aggregated utility of all activity participants in a series of test scenarios. **Therefore, this method can be used as-is in practice without the need of any modification**.

In Chapter 5 was developed a sequential heuristic search algorithm for changing departure times of public transport modes in near real time (in a matter of minutes) for adjusting to the arrival time needs of joint leisure activity passengers without deteriorating the KPIs of public transport operations. The sequential heuristic search was validated after the implementation in a test-case focusing on bus lines 1 and 4 in Stockholm and Twitter data for deriving individual trips from 62 users to joint leisure activity locations. Due to the schedule changes after optimization, the operational performance of bus services demonstrated an EWT improvement at a service-wide level for all services while only some stations from line 1, direction 2 and line 4, direction 4 had a slight EWT deterioration(up to 0.6min.); without affecting significantly though the level of service of bus operations. At the same time, after the re-scheduling, the joint leisure activity trips from all services enjoyed new arrival times to the joint activity stations which were more than 50% closer to the starting times of those activities. During the validation stage, it was not possible to define how close was the sequential heuristic search solution to the global optimum given the computational complexity of the problem that hinders the computation of the global optimum in the first place. However, it is guaranteed that the solutions of the heuristic sequential search always improve the current operations due to the greedy nature of the search and **the method can be used as-is in practical applications**.

Finally, two methods for finding the shortest multi-modal path for a user who is willing to travel from his/her current location to the location of a joint leisure activity were developed in Chapter 6. Those methods considered the specific preferences of each user and were based on an iterative label-setting approach where each label represented a distinct path. The first method was an exact label-setting method that always converged to the global optimum and was tested for validation purposes in simulated networks with up to 3,000 transport nodes and 30,000 transport links. However, it was discovered during the validation phase that this method had to generate a vast amount of labels prior to its convergence and that slowed down its implementation especially in the case of multiple user-preference constraints. Therefore, this method **can be used as-is in practical applications at small to medium-sized networks**. For resolving the computational scalability problem, the shortest paths from all locations of the network to the location of the joint leisure activity were pre-computed using historical travel time data and a second method based on an aggressive label-pruning strategy was developed. This method had a heuristic nature and converged closer to the global optimum when the estimated average travel times were closer to the real ones. In practical implementations at simulated networks that considered up to 4 user preference constraints, the aggressive label-pruning method outperformed vastly the exact label-setting method in terms of operational costs by converging in a matter of seconds. This enables **its practical implementation as-is in real-world route and mode-selection applications** with an accuracy error of less than 10% according with the validation tests.

8.3. Future Work

This thesis provided a comprehensive approach for optimizing trips related to non-recurrent activities in urban environments. This approach depends on the data input due to its high granularity. Especially the need of user-generated geo-location data is an application barrier; however, it is expected to overcome this barrier as more and more transport authorities open up their data and SmartCard logs become publicly available on the web including tap-ins and tap-outs of passengers. In this direction, the fusion of user-generated geo-location data from social media, smartcards, cellular cells and PDAs will be of paramount importance for improving the accuracy of computational learning of users' mobility/activity patterns and is recommended as future work topic.

In addition, the utility-maximization model of capturing users' willingness of traveling certain distances to participate at different activity types can be coupled with zonal models that provide insights on the activity options of different zones. For instance, if there are two similar leisure activity options, an individual might be willing to travel more and participate at the one that is far away from his/her current location if that activity is in a zone that offers several other activity options (zone with more attractiveness). Studying only geo-location datasets from users cannot reveal that information; therefore, research on the attractiveness of zones and their importance on leisure activity participation will be beneficial.

In future research, experiments on other public transport modes can also be conducted for developing demand-responsive schemes for covering joint leisure activity trips (i.e., train, underground, tram).

A final challenge is to develop the required applications that will improve the usability of the comprehensive non-recurrent activity optimization approached described in this thesis consisting of:

- a smartphone application where joint leisure activity *location* and *time* suggestions are provided to users based on their user-generated geo-location data logs

- a smartphone application where users are informed about changes on demand-responsive public transportation for increasing the service competitiveness of public transportation
- a smartphone application which proposes the optimal path to the location of the activity considering multiple modes and users' preferences in the form of resource constraints

Bibliography

References in citation order.

[1] *Transport for London, Travel in London, London Travel Demand Survey*. http://www.tfl.gov.uk/cdn/static/cms/documents/london-travel-demand-survey.pdf. Accessed: 2014-09-30. 2014 (cited on pages 5, 45).

[2] *New York Regional Travel Survey*. http://www.nymtc.org/project/surveys/survey.html. Accessed: 2014-09-30. 2014 (cited on pages 5, 45).

[3] Marta C. Gonzàlez, Cèsar A. Hidalgo, and Albert-Làszlò Barabàsi. 'Understanding individual human mobility patterns'. en. In: *Nature* 453.7196 (June 2008), pp. 779–782. (Visited on 06/21/2013) (cited on pages 6, 8, 27).

[4] B. Ferris, K. Watkins, and A. Borning. 'OneBusAway: Results from Providing Real-Time Arrival Information for Public Transit'. In: *In Proc. of the 28th Intl. Conf. on Human Factors in Computing Systems (CHI '10)* (June 2010) (cited on page 7).

[5] Kay W. Axhausen and Tommy Gärling. 'Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems'. In: *Transport Reviews* 12.4 (1992), pp. 323–341. (Visited on 04/15/2013) (cited on pages 7, 27).

[6] Theo A Arentze and Harry J.P Timmermans. 'A learning-based transportation oriented simulation system'. In: *Transportation Research Part B: Methodological* 38.7 (Aug. 2004), pp. 613–633. (Visited on 04/15/2013) (cited on pages 7, 27).

[7] Ram Pendyala et al. 'Florida Activity Mobility Simulator: Overview and Preliminary Validation Results'. In: *Transportation Research Record: Journal of the Transportation Research Board* 1921.-1 (Jan. 2005), pp. 123–130. (Visited on 04/15/2013) (cited on pages 7, 27).

[8] Dung-Ying Lin et al. 'Integration of Activity-Based Modeling and Dynamic Traffic Assignment'. In: *Transportation Research Record: Journal of the Transportation Research Board* 2076.-1 (Dec. 2008), pp. 52–61. (Visited on 04/15/2013) (cited on pages 7, 27).

[9] Mirco Musolesi and Cecilia Mascolo. 'Designing Mobility Models Based on Social Network Theory'. In: *SIGMOBILE Mob. Comput. Commun. Rev.* 11.3 (July 2007), pp. 59–70. (Visited on 05/21/2014) (cited on pages 7, 27).

[10] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. 'Interdependence and predictability of human mobility and social interactions'. In: *Pervasive and Mobile Computing* 9.6 (2013), pp. 798–807 (cited on page 7).

[11] Juan Antonio Carrasco et al. 'Collecting social network data to study social activity-travel behavior: an egocentric approach'. In: *Environment and Planning B: Planning and Design* 35.6 (2008), pp. 961–980. (Visited on 04/10/2013) (cited on page 7).

[12] Theo Arentze and Harry Timmermans. 'Social networks, social interactions, and activity-travel behavior: a framework for microsimulation'. In: *Environment and planning. B, Planning & design* 35.6 (2008), p. 1012 (cited on page 7).

[13] Ying Chen, Andreas Frei, and Hani S Mahmassani. 'From Personal Attitudes to Public Opinion: Information Diffusion in Social Networks Toward Sustainable Transportation'. In: *Transportation Research Board 93rd Annual Meeting*. 14-3566. 2014 (cited on page 7).

[14] Francesco Calabrese et al. 'Estimating Origin-Destination Flows Using Mobile Phone Location Data'. In: *IEEE Pervasive Computing* 10.4 (2011), pp. 36–44 (cited on pages 8, 27).

[15] F. Calabrese et al. 'Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome'. In: *IEEE Transactions on Intelligent Transportation Systems* 12.1 (Mar. 2011), pp. 141–151 (cited on pages 8, 27).

[16] Richard Becker et al. 'Human mobility characterization from cellular network data'. In: *Communications of the ACM* 56.1 (2013), pp. 74–82 (cited on page 8).

[17] J. White and I. Wells. 'Extracting origin destination information from mobile phone data'. In: *Road Transport Information and Control, 2002. Eleventh International Conference on (Conf. Publ. No. 486)*. 2002, pp. 30–34 (cited on pages 8, 27).

[18] Yi Zhang et al. 'Daily O-D Matrix Estimation Using Cellular Probe Data'. In: *Transportation Research Board: 89th Annual Meeting* (2010). (Visited on 11/06/2013) (cited on pages 8, 27).

[19] Changxuan Pan et al. 'Cellular-Based Data-Extracting Method for Trip Distribution'. In: *Transportation Research Record: Journal of the Transportation Research Board* 1945.-1 (Jan. 2006), pp. 33–39. (Visited on 11/06/2013) (cited on pages 8, 27).

[20] Goran M Djuknic and Robert E Richton. 'Geolocation and assisted GPS'. In: *Computer* 34.2 (2001), pp. 123–125 (cited on page 8).

[21] Keemin Sohn and Daehyun Kim. 'Dynamic origin–destination flow estimation using cellular communication system'. In: *Vehicular Technology, IEEE Transactions on* 57.5 (2008), pp. 2703–2713 (cited on page 8).

[22] Francesco Calabrese et al. 'Understanding individual mobility patterns from urban sensing data: A mobile phone trace example'. In: *Transportation research part C: emerging technologies* 26 (2013), pp. 301–313 (cited on page 8).

[23] C. Song et al. 'Limits of Predictability in Human Mobility'. In: *Science* 327.5968 (2010), pp. 1018–1021 (cited on page 8).

[24] Honghui Dong et al. 'Urban Traffic Commuting Analysis Based on Mobile Phone Data'. In: *in Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems* (2014), pp. 605–610 (cited on page 8).

[25] Mingdao Wu et al. 'Traffic Semantic Analysis Based on Mobile Phone Base Station Data'. In: *in Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems* (2014), pp. 611–616 (cited on page 8).

[26] Hiroki Ohasi et al. 'Trip-separation method using sensor data continuously collected by smartphone'. In: *in Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems* (2014), pp. 2966–2972 (cited on page 8).

[27] Hiroki Ohashi et al. 'Modality Classification Method Based on the Model of Vibration Generation while Vehicles are Running'. In: *Proceedings of the Sixth ACM SIGSPATIAL International Workshop on Computational Transportation Science*. ACM. 2013, p. 37 (cited on page 8).

[28] Pan Hui et al. 'Pocket Switched Networks and Human Mobility in Conference Environments'. In: *Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-tolerant Networking*. WDTN '05. New York, NY, USA: ACM, 2005, pp. 244–251 (cited on pages 8, 27).

[29] Augustin Chaintreau et al. *Pocket switched networks: Real-world mobility and its consequences for opportunistic forwarding*. Tech. rep. Technical Report UCAM-CL-TR-617, University of Cambridge, Computer Laboratory, 2005 (cited on pages 8, 27).

[30] Satish V Ukkusuri et al. 'Exploring Crisis Informatics Using Social Media Data: A Study on 2013 Oklahoma Tornado'. In: *Transportation Research Board 93rd Annual Meeting*. 14-2099. 2014 (cited on page 9).

[31] Sarah Vieweg et al. 'Collective intelligence in disaster: An examination of the phenomenon in the aftermath of the 2007 Virginia Tech shootings'. In: *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM)*. 2008 (cited on page 9).

[32] Amanda L Hughes et al. 'Sight-Seeing in Disaster: An Examination of On-Line Social Convergence'. In: *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM)*. 2008 (cited on page 9).

[33] Yan Qu, Philip Fei Wu, and Xiaoqing Wang. 'Online community response to major disaster: A study of Tianya forum in the 2008 Sichuan earthquake'. In: *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*. IEEE. 2009, pp. 1–11 (cited on page 9).

[34] Yan Qu et al. 'Microblogging after a major disaster in China: a case study of the 2010 Yushu earthquake'. In: *Proceedings of the ACM 2011 conference on Computer supported cooperative work*. ACM. 2011, pp. 25–34 (cited on page 9).

[35] Sarah Vieweg et al. 'Microblogging during two natural hazards events: what twitter may contribute to situational awareness'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2010, pp. 1079–1088 (cited on page 9).

[36] Francisco C Pereira, Filipe Rodrigues, and Moshe Ben-Akiva. 'Using data from the web to predict public transport arrivals under special events scenarios'. In: *Journal of Intelligent Transportation Systems* just-accepted (2013) (cited on page 9).

[37] K. Gkiotsalitis and A. Alexandrou. 'Mobility Demand Prediction in Urban Scenarios through Multi-source, User-generated Data'. In: *18th Pan-American Conference of Traffic and Transportation Engineering and Logistics, PANAM 2014* (June 2014) (cited on pages 9, 11).

[38] Konstantinos Gkiotsalitis, Francesco Alesiani, and Roberto Baldessari. 'Educated Rules for the Prediction of Human Mobility Patterns based on Sparse Social Media and Mobile Phone Data'. In: *Transportation Research Board: 93rd Annual Meeting* (2014), pp. 14–0745. (Visited on 11/06/2013) (cited on page 9).

[39] Francesco Alesiani, Konstantinos Gkiotsalitis, and Roberto Baldessari. 'A Probabilistic Activity Model for Predicting the Mobility Patterns of Homogeneous Social Groups based on Social Network Data'. In: *Transportation Research Board: 93rd Annual Meeting* (2014), pp. 14–1013. (Visited on 11/06/2013) (cited on page 10).

[40] Lorenzo Gabrielli et al. 'From tweets to semantic trajectories: mining anomalous urban mobility patterns'. In: *Citizen in Sensor Networks*. Springer, 2014, pp. 26–35 (cited on page 10).

[41] Anastasios Noulas et al. 'An Empirical Study of Geographic User Activity Patterns in Foursquare.' In: *ICwSM* 11 (2011), pp. 70–573 (cited on page 10).

[42] Anastasios Noulas et al. 'A tale of many cities: universal patterns in human urban mobility'. In: *PloS one* 7.5 (2012), e37027 (cited on page 10).

[43] Bartosz Hawelka et al. 'Geo-located Twitter as proxy for global mobility patterns'. In: *Cartography and Geographic Information Science* 41.3 (2014), pp. 260–271 (cited on page 10).

[44] Samiul Hasan, Xianyuan Zhan, and Satish V Ukkusuri. 'Understanding urban human activity and mobility patterns using large-scale location-based data from online social media'. In: *Proceedings of the 2nd ACM SIGKDD international workshop on urban computing*. ACM. 2013, p. 6 (cited on page 10).

[45] M.P. Pelletier, M. Trpanier, and C. Morency. 'Smart card data use in public transit: A literature review'. In: *Transportation Research Part C: Emerging Technologies* 19.4 (2011), pp. 557–568 (cited on page 10).

[46] Nigel HM Wilson, Jinhua Zhao, and Adam Rahbee. 'The potential impact of automated data collection systems on urban public transport planning'. In: *Schedule-Based Modeling of Transportation Networks* 46 (2009), pp. 1–25 (cited on page 10).

[47] A. Chatterjee and M.M. Venigalla. *Travel Demand Forecasting for Urban Transportation Planning*. en. In Handbook of Transportation Engineering, Volume I: Systems and Operations, 2011 (cited on page 10).

[48] C. Smith, D. Quercia, and L. Capra. 'Anti-gravity underground?' In: *In Proceedings of the Second Workshop on Pervasive Urban Applications*. PURBA '12, 2012 (cited on page 10).

[49] Irina Ceapa, Chris Smith, and Licia Capra. 'Avoiding the crowds: understanding tube station congestion patterns from trip data'. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. ACM. 2012, pp. 134–141 (cited on page 10).

[50] Catherine Morency, Martin Trepanier, and Bruno Agard. 'Measuring transit use variability with smart-card data'. In: *Transport Policy* 14.3 (2007), pp. 193–203 (cited on page 10).

[51] Wonjae Jang. 'Travel time and transfer analysis using transit smart card data'. In: *Transportation Research Record: Journal of the Transportation Research Board* 2144.1 (2010), pp. 142–149 (cited on page 10).

[52] Mousumi Bagchi and PR White. 'The potential of public transport smart card data'. In: *Transport Policy* 12.5 (2005), pp. 464–474 (cited on page 10).

[53] Chen Zhong et al. 'Measuring variability of mobility patterns from multiday smart-card data'. In: *Journal of Computational Science* 9 (2015), pp. 125–130 (cited on page 10).

[54] Stefan Foell et al. 'Catch me if you can: Predicting Mobility Patterns of Public Transport Users'. In: *in Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems* (2014), pp. 1983–1990 (cited on page 11).

[55] Jordan Ivanchev, Heiko Aydt, and Alois Knoll. 'Stochastic Bus Traffic Modelling and Validation Using Smart Card Fare collection data'. In: *in Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems* (2014), pp. 1983–1990 (cited on page 11).

[56] Daming Li et al. 'Estimating a transit passenger trip origin-destination matrix using automatic fare collection system'. In: *International Conference on Database Systems for Advanced Applications*. Springer. 2011, pp. 502–513 (cited on page 11).

[57] Marcela A Munizaga and Carolina Palma. 'Estimation of a disaggregate multimodal public transport Origin–Destination matrix from passive smartcard data from Santiago, Chile'. In: *Transportation Research Part C: Emerging Technologies* 24 (2012), pp. 9–18 (cited on page 11).

[58] Artem Chakirov and Alexander Erath. 'Use of public transport smart card fare payment data for travel behaviour analysis in Singapore'. In: (2011) (cited on page 11).

[59] John Krumm and Eric Horvitz. 'Predestination: Where do you want to go today?' In: *Computer* 40.4 (2007) (cited on page 11).

[60] Lin Liao et al. 'Building personal maps from GPS data'. In: *Annals of the New York Academy of Sciences* 1093.1 (2006), pp. 249–265 (cited on page 11).

[61] Lin Liao et al. 'Learning and inferring transportation routines'. In: *Artificial Intelligence* 171.5-6 (2007), pp. 311–331 (cited on page 11).

[62] Jun Wu et al. 'Automated time activity classification based on global positioning system (GPS) tracking data'. In: *Environ Health* 10 (2011), p. 101 (cited on page 11).

[63] Eiji Hato. 'Development of behavioral context addressable loggers in the shell for travel-activity analysis'. In: *Transportation Research Part C: Emerging Technologies* 18.1 (2010), pp. 55–67 (cited on page 11).

[64] Quannan Li et al. 'Mining user similarity based on location history'. In: *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*. ACM. 2008, p. 34 (cited on pages 11, 12).

[65] Wendy Bohte, Kees Maat, and Wilko Quak. 'A method for deriving trip destinations and modes for GPS-based travel surveys'. In: *Research in Urbanism Series* 1.1 (2008), pp. 127–143 (cited on page 11).

[66] Cynthia Chen et al. 'Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study'. In: *Transportation Research Part A: Policy and Practice* 44.10 (2010), pp. 830–840 (cited on page 11).

[67] Ming Li et al. 'Trip analyzer through smartphone apps'. In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM. 2011, pp. 537–540 (cited on page 12).

[68] Apichon Witayangkurn et al. 'Trip Reconstruction and Transportation Mode Extraction on Low Data Rate GPS Data from Mobile Phone'. In: *in Proceedings of the International Conference on Computers in Urban Planning and Urban Management(CUPUM 2013)* (2013) (cited on page 12).

[69] Francisco C. Pereira, Filipe Rodrigues, and Moshe Ben-Akiva. 'Using data from the web to predict public transport arrivals under special events scenarios'. In: *Journal of Intelligent Transportation Systems* 0.ja (Jan. 2014l), null. (Visited on 05/21/2014) (cited on page 27).

[70] Wei Xiong Shang et al. 'Predicting dynamic transportation demand with mobility data'. US20120191505 A1. International Classification: G06Q10/04 U.S. Classification: H04W 4/02P6; G06Q 30/0202. July 2012 (cited on page 27).

[71] Ernst K Berndt et al. 'Estimation and inference in nonlinear structural models'. In: *Annals of Economic and Social Measurement, Volume 3, number 4*. NBER, 1974, pp. 103–116 (cited on page 36).

[72] Martin Ester et al. 'A density-based algorithm for discovering clusters in large spatial databases with noise'. In: AAAI Press, 1996, pp. 226–231 (cited on page 37).

[73] Gonçalo Homem de Almeida Correia and António Pais Antunes. 'Optimization approach to depot location and trip selection in one-way carsharing systems'. In: *Transportation Research Part E: Logistics and Transportation Review* 48.1 (2012), pp. 233–247 (cited on page 59).

[74] Niels Agatz et al. 'Optimization for dynamic ride-sharing: A review'. In: *European Journal of Operational Research* 223.2 (2012), pp. 295–303 (cited on page 59).

[75] Eric J Miller, Matthew J Roorda, and Juan Antonio Carrasco. 'A tour-based model of travel mode choice'. In: *Transportation* 32.4 (2005), pp. 399–422 (cited on page 63).

[76] Javier Asensio. 'Transport mode choice by commuters to Barcelona's CBD'. In: *Urban Studies* 39.10 (2002), pp. 1881–1895 (cited on page 63).

[77] John L Bowman and Moshe E Ben-Akiva. 'Activity-based disaggregate travel demand model system with activity schedules'. In: *Transportation Research Part A: Policy and Practice* 35.1 (2001), pp. 1–28 (cited on page 63).

[78] Geraldo Regis Mauri, Luiz Antonio, and Nogueira Lorena. 'Customers' satisfaction in a dial-a-ride problem'. In: *IEEE Intelligent Transportation Systems Magazine* 1.3 (2009), pp. 6–14 (cited on page 63).

[79] Elodie Castex, S Houzet, and D Josselin. 'Prospective research in the technological and mobile society: new Demand Responsive Transports for new territories to serve'. In: *Proceeding of the 7th AGILE Conference, Heraklion*. Vol. 29. 2004, pp. 559–568 (cited on page 63).

[80] Janet E Dickinson et al. 'Tourism and the smartphone app: Capabilities, emerging practice and scope in the travel domain'. In: *Current Issues in Tourism* 17.1 (2014), pp. 84–101 (cited on page 64).

[81] Shlomo Bekhor, Moshe E Ben-Akiva, and M Scott Ramming. 'Evaluation of choice set generation algorithms for route choice models'. In: *Annals of Operations Research* 144.1 (2006), pp. 235–247 (cited on page 83).

[82] Jennifer Margaret Farver. 'Hybrid vehicle-centric route guidance'. PhD thesis. Massachusetts Institute of Technology, 2005 (cited on page 85).

[83] David E Kaufman, Robert L Smith, and Karl E Wunderlich. 'An iterative routing/assignment method for anticipatory real-time route guidance'. In: *Vehicle Navigation and Information Systems Conference, 1991*. Vol. 2. IEEE. 1991, pp. 693–700 (cited on page 85).

[84] Moshe Ben-Akiva et al. 'DynaMIT: a simulation-based system for traffic prediction'. In: *DACCORS short term forecasting workshop, The Netherlands*. Citeseer. 1998 (cited on page 85).

[85] Moshe Ben-Akiva, Andre de Palma, and Isam Kaysi. 'The impact of predictive information on guidance efficiency: An analytical approach'. In: *Advanced methods in transportation analysis*. Springer, 1996, pp. 413–432 (cited on page 85).

[86] J Bottom et al. 'Investigation of route guidance generation issues by simulation with DynaMIT'. In: *International symposium Pergamon on transportation and traffic theory*. Citeseer. 1999 (cited on page 85).

[87] Jon Bottom et al. 'Generation of consistent anticipatory route guidance'. In: *Proceedings of TRISTAN III, San Juan, Puerto Rico* 2 (1998) (cited on page 86).

[88] Alexander Sohr and Peter Wagner. 'Project drive test–comparison of five Personal Navigation Devices (PND)'. In: (2011) (cited on page 86).

[89] Andrew V Goldberg, Haim Kaplan, and Renato F Werneck. 'Reach for a: Efficient point-to-point shortest path algorithms'. In: *Proceedings of the Meeting on Algorithm Engineering & Experiments*. Society for Industrial and Applied Mathematics. 2006, pp. 129–143 (cited on page 88).

[90] Dominik Schultes. 'Fast and exact shortest path queries using highway hierarchies'. In: *Master-Arbeit, Universität des Saarlandes, Saarbrücken* (2005) (cited on page 88).

[91] Hsun-Jung Cho and Chien-Lun Lan. 'Hybrid shortest path algorithm for vehicle navigation'. In: *The Journal of Supercomputing* 49.2 (2009), pp. 234–247 (cited on page 88).

[92] Liping Fu and Larry R Rilett. 'Expected shortest paths in dynamic and stochastic traffic networks'. In: *Transportation Research Part B: Methodological* 32.7 (1998), pp. 499–516 (cited on page 89).

[93] Hesham Ahmed Rakha et al. 'Estimating path travel-time reliability'. In: *2006 IEEE Intelligent Transportation Systems Conference*. IEEE. 2006, pp. 236–241 (cited on page 89).

[94] I Kaparias and MGH Bell. 'Testing a reliable in-vehicle navigation algorithm in the field'. In: *IET Intelligent Transport Systems* 3.3 (2009), pp. 314–324 (cited on page 89).

[95] Mohamed A Abdel-Aty, Ryuichi Kitamura, and Paul P Jovanis. 'Using stated preference data for studying the effect of advanced traffic information on drivers' route choice'. In: *Transportation Research Part C: Emerging Technologies* 5.1 (1997), pp. 39–50 (cited on page 89).

[96] Ting-Yu Chen, Hsin-Li Chang, and Gwo-Hshiung Tzeng. 'Using a weight-assessing model to identify route choice criteria and information effects'. In: *Transportation Research Part A: Policy and Practice* 35.3 (2001), pp. 197–224 (cited on page 89).

[97] Kenneth Train and Wesley W Wilson. 'Estimation on stated-preference experiments constructed from revealed-preference choices'. In: *Transportation Research Part B: Methodological* 42.3 (2008), pp. 191–203 (cited on page 89).

[98] Maria P Boile. 'Intermodal transportation network analysis-A GIS application'. In: *Electrotechnical Conference, 2000. MELECON 2000. 10th Mediterranean*. Vol. 2. IEEE. 2000, pp. 660–663 (cited on page 90).

[99] Zhuo Sun et al. 'A user preferable k-shortest path algorithm for intermodal network'. In: *Proceedings of the Eastern Asia Society for Transportation Studies*. Vol. 2007. 0. Eastern Asia Society for Transportation Studies. 2007, pp. 240–240 (cited on page 90).

[100] Dimitri P Bertsekas. 'A simple and fast label correcting algorithm for shortest paths'. In: *Networks* 23.8 (1993), pp. 703–709 (cited on page 90).

[101] Athanasios Ziliaskopoulos and Whitney Wardell. 'An intermodal optimum path algorithm for multimodal networks with dynamic arc travel times and switching delays'. In: *European Journal of Operational Research* 125.3 (2000), pp. 486–502 (cited on page 90).

[102] Angelica Lozano and Giovanni Storchi. 'Shortest viable path algorithm in multimodal networks'. In: *Transportation Research Part A: Policy and Practice* 35.3 (2001), pp. 225–241 (cited on page 90).

[103] Mounir Boussedjra, Christelle Bloch, and Abdellah El Moudni. 'Solution optimale pour la recherche du meilleur chemin intermodal'. In: *Roc. 4e Conference Francophone de Modelisation et SIMulation MOSIM*. 2003, pp. 230–237 (cited on page 90).

[104] Bonnie S Boardman et al. 'Computer assisted routing of intermodal shipments'. In: *Computers & Industrial Engineering* 33.1 (1997), pp. 311–314 (cited on page 90).

[105] Paola Modesti and Anna Sciomachen. 'A utility measure for finding multiobjective shortest paths in urban multimodal transportation networks'. In: *European Journal of Operational Research* 111.3 (1998), pp. 495–508 (cited on page 91).

[106] Jeffrey M Jaffe. 'Algorithms for finding paths with multiple constraints'. In: *Networks* 14.1 (1984), pp. 95–116 (cited on page 91).

[107] Moshe Dror. 'Note on the complexity of the shortest path models for column generation in VRPTW'. In: *Operations Research* 42.5 (1994), pp. 977–978 (cited on page 91).

[108] Martin Desrochers and François Soumis. 'A reoptimization algorithm for the shortest path problem with time windows'. In: *European Journal of Operational Research* 35.2 (1988), pp. 242–254 (cited on page 91).

[109] Dominique Feillet et al. 'An exact algorithm for the elementary shortest path problem with resource constraints: Application to some vehicle routing problems'. In: *Networks* 44.3 (2004), pp. 216–229 (cited on page 91).

[110] Stefan Irnich and Guy Desaulniers. 'Shortest path problems with resource constraints'. In: *Column generation*. Springer, 2005, pp. 33–65 (cited on page 91).

[111] Giovanni Righini and Matteo Salani. 'New dynamic programming algorithms for the resource constrained elementary shortest path problem'. In: *Networks* 51.3 (2008), pp. 155–170 (cited on page 91).

[112] Gabriel Y Handler and Israel Zang. 'A dual algorithm for the constrained shortest path problem'. In: *Networks* 10.4 (1980), pp. 293–309 (cited on page 91).

[113] John E Beasley and Nicos Christofides. 'An algorithm for the resource constrained shortest path problem'. In: *Networks* 19.4 (1989), pp. 379–394 (cited on page 91).

[114] Kurt Mehlhorn and Mark Ziegelmann. 'Resource constrained shortest paths'. In: *European Symposium on Algorithms*. Springer. 2000, pp. 326–337 (cited on page 91).

[115] Irina Dumitrescu and Natashia Boland. 'Improved preprocessing, labeling and scaling algorithms for the weight-constrained shortest path problem'. In: *Networks* 42.3 (2003), pp. 135–153 (cited on page 91).

[116] Pasquale Avella, Maurizio Boccia, and Antonio Sforza. 'A penalty function heuristic for the resource constrained shortest path problem'. In: *European Journal of Operational Research* 142.2 (2002), pp. 221–230 (cited on page 91).

[117] Anass Nagih and François Soumis. 'Nodal aggregation of resource constraints in a shortest path problem'. In: *European Journal of Operational Research* 172.2 (2006), pp. 500–514 (cited on page 92).

[118] Xiaoyan Zhu and Wilbert E Wilhelm. 'Three-stage approaches for optimizing some variations of the resource constrained shortest-path sub-problem in a column generation context'. In: *European journal of operational research* 183.2 (2007), pp. 564–577 (cited on page 92).

[119] Paolo Serafini. 'Some considerations about computational complexity for multi objective combinatorial problems'. In: *Recent advances and historical development of vector optimization*. Springer, 1987, pp. 222–232 (cited on page 92).

[120] Chi Tung Tung and Kim Lin Chew. 'A bicriterion Pareto-optimal path algorithm.' In: *ASIA-PACIFIC J. OPER. RES.* 5.2 (1988), pp. 166–172 (cited on page 92).

[121] Chi Tung Tung and Kim Lin Chew. 'A multicriteria Pareto-optimal path algorithm'. In: *European Journal of Operational Research* 62.2 (1992), pp. 203–209 (cited on page 92).

[122] Ernesto de Queros Vieira Martins and JLE Santos. 'The labeling algorithm for the multiobjective shortest path problem'. In: *Departamento de Matematica, Universidade de Coimbra, Portugal, Tech. Rep. TR-99/005* (1999) (cited on page 92).

[123] José Manuel Paixão and José Luis Santos. 'Labeling Methods for the General Case of the Multi-objective Shortest Path Problem– A Computational Study'. In: *Computational Intelligence and Decision Making*. Springer, 2013, pp. 489–502 (cited on page 92).

[124] Anders JV Skriver and Kim Allan Andersen. 'A label correcting approach for solving bicriterion shortest-path problems'. In: *Computers & Operations Research* 27.6 (2000), pp. 507–524 (cited on page 92).

[125] VN Sastry, TN Janakiraman, and S Ismail Mohideen. 'New algorithms for multi objective shortest path problem'. In: *Opsearch* 40.4 (2003), pp. 278–298 (cited on page 92).

[126] Andrea Raith and Matthias Ehrgott. 'A comparison of solution strategies for biobjective shortest path problems'. In: *Computers & Operations Research* 36.4 (2009), pp. 1299–1331 (cited on page 92).

[127] Michael L Fredman and Robert Endre Tarjan. 'Fibonacci heaps and their uses in improved network optimization algorithms'. In: *Journal of the ACM (JACM)* 34.3 (1987), pp. 596–615 (cited on page 100).

APPENDIX

Appendix A.

A.1. Curriculum Vitae

Personal Data

- Name: Konstantinos Gkiotsalitis
- Address: Pfannmüllerstrasse 28, 60488 Frankfurt am Main
- Occupation: Research Scientist at NEC Laboratories Europe (2015-2017) | Research Associate at NEC Laboratories Europe (2012-2015)

Education

- 2013-2017: PhD Candidate at the National Technical University of Athens
- 2011-2012: Intercollegiate MSc in Transport and Sustainable Development at Imperial College London and University College London
- 2005-2010: 5-year diploma in Civil Engineering from the National Technical University of Athens (specialization at the Department of Transportation Planning and Engineering)
- *Period 2010-2011: Greek Army Service

Peer-Reviewed Journal Publications

1. K. Gkiotsalitis, A. Stathopoulos (2015): "A mobile application for real-time multimodal routing under a set of users' preferences", Journal of Intelligent Transportation Systems 19 (2), 149-166.
2. K. Gkiotsalitis, A. Stathopoulos (2015): "A utility-maximization model for retrieving users' willingness to travel for participating in activities from big-data", Transportation Research Part C: Emerging Technologies 58, 265-277.
3. K Gkiotsalitis, AHF Chow (2014): "Significance of fundamental diagrams to first-order macroscopic traffic modelling", International Journal of Transportation 2 (2), 15-32.
4. AHF Chow, Y Li, K Gkiotsalitis (2015): "Specifications of fundamental diagrams for dynamic traffic modeling", Journal of Transportation Engineering 141 (9), 04015015.
5. K. Gkiotsalitis, A. Stathopoulos (2015): "Optimizing Leisure Travel: Is BigData Ready to Improve the Joint Leisure Activities Efficiency?", vol.38 of the series Engineering and Applied Sciences Optimization, Springer International Publishing, 53-71.
6. K. Gkiotsalitis, A. Stathopoulos (2016): "Joint leisure travel optimization with user-generated data via perceived utility maximization", Transportation Research Part C: Emerging Technologies 68, 532-548.

7. K. Gkiotsalitis, A. Stathopoulos (2016): "Demand-Responsive Public Transportation Re-scheduling for adjusting to the Joint Leisure Activity Demand", International Journal of Transportation Science and Technology (IJTST), Elsevier, vol. 5, issue 2, 68-82.
8. K. Gkiotsalitis, N. Maslekar (2016): "Dynamic Bus Operations Optimization with REFLEX", NEC Journal of Artificial Intelligence, vol. 69, no 1, 68-72.
9. K. Gkiotsalitis, N. Maslekar (2017): "Towards transfer synchronization of regularity-based bus operations with sequential hill-climbing", to appear at the Journal of Public Transport, Springer.

Conference Publications

1. F. Alesiani, K. Gkiotsalitis, R. Baldessari (2014): "A probabilistic activity model for predicting the mobility patterns of homogeneous social groups based on social network data", Transportation Research Board, 93rd Annual Meeting.
2. K. Gkiotsalitis, A. Alexandrou (2014): "Mobility demand prediction in urban scenarios through multi-source, user-generated data", 18th Pan-American conference of traffic and transportation engineering (PANAM).
3. K. Gkiotsalitis, F. Alesiani, R. Baldessari (2014): "Educated rules for the prediction of human mobility patterns based on sparse social media and mobile phone data", Transportation Research Board, 93rd Annual Meeting.
4. K. Gkiotsalitis, F. Alesiani (2014): "Opportunistic solution-space reduction techniques for reducing the time complexity of Dynamic Speed Control with microsimulation on motorways", 17th International IEEE Conference on Intelligent Transport Systems (ITSC), 1788-1795.
5. K. Gkiotsalitis, N. Maslekar (2015): "Improving Bus Service Reliability with Stochastic Optimization", 18th International IEEE Conference on Intelligent Transport Systems (ITSC), 2794-2799.
6. K. Gkiotsalitis, N. Maslekar (2016): "Sequential Evolutionary optimization for improving Regularity-based Bus Services", Transportation Research Board, 96th Annual Meeting.
7. K. Gkiotsalitis, O. Cats (2016): "Exact optimization of Bus Frequency Settings considering Demand and Trip time variations", Transportation Research Board, 96th Annual Meeting.

Patents

1. K. Gkiotsalitis, N. Maslekar (2016): "Method for providing configuration information for a system comprising a plurality of moving objects", WO Patent WO/2016/155, 790
2. K Gkiotsalitis, F. Alesiani (2015): "Method and computer program product for accurate motorway speed control", WO Patent WO2015049064 A1
3. K. Gkiotsalitis, N. Maslekar (2015): "Dynamic Fleet Routing", WO Patent WO/2015/154, 831

Awards

- NEC Laboratories Award for Smart Public Transportation Demonstration at the NEC OpenHouse 2015 in Tokyo
- NEC Laboratories Award for the Highway Speed Control Demonstrator at the 2013 ITS World Congress in Tokyo
- OIKOPOLIS Award for the best dissertation of the year in Greece (June 2011) for pioneering in the field of transport and sustainable development under the supervision of Prof. Antony Stathopoulos

Reviewer

- IEEE Intelligent Transport Systems Conference (ITSC)
- Transportation Research Board

Organizations

- Chartered Institution of Highways and Transportation (MCIHT: 000072011). Serving as Overseas Champion for Germany since May 2012
- Technical Chamber of Greece (TEE)
- SES (Hellenic Institute of Transportation Engineers)
- Imperial College London Alumni (00706114)

Foreign Languages

- English
- German