Project Acronym: OPTIMUM
Grant Agreement number: 636160 (H2020-MG-7.1 - RIA)
Project Full Title: Multi-source Big Data Fusion Driven Proactivity for Intelligent Mobility

DELIVERABLE

**D3.1 – Orientation and Forecasting Blueprint**

<table>
<thead>
<tr>
<th>Dissemination level</th>
<th>PU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of Document</td>
<td>Report</td>
</tr>
<tr>
<td>Contractual date of delivery</td>
<td>M12 – 30-04-2016</td>
</tr>
<tr>
<td>Actual Date of Delivery</td>
<td>M12 – 30-04-2016</td>
</tr>
<tr>
<td>Deliverable Number</td>
<td>D3.1</td>
</tr>
<tr>
<td>Deliverable Name</td>
<td>Orientation and Forecasting Blueprint</td>
</tr>
<tr>
<td>Deliverable Leader</td>
<td>UAEGEAN</td>
</tr>
<tr>
<td>Work package(s)</td>
<td>WP3</td>
</tr>
<tr>
<td>Status &amp; version</td>
<td>1.0 – Final</td>
</tr>
<tr>
<td>Number of pages</td>
<td>68</td>
</tr>
<tr>
<td>WP contributing to the deliverable</td>
<td>WP3</td>
</tr>
<tr>
<td>WP / Task responsible</td>
<td>WP3 / T1.3</td>
</tr>
<tr>
<td>Coordinator (name / contact)</td>
<td>Mr. Konstantinos Thiveos (INTRASOFT International)</td>
</tr>
<tr>
<td>Other Contributors</td>
<td>JSI, NISSA, ICCS, UoW</td>
</tr>
<tr>
<td>EC Project Officer</td>
<td>Mr. Walter Mauritsch (EC, INEA)</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Orientation, forecasting, blueprint, behavioural, traffic prediction, complex events analysis, model, mobility patterns.</td>
</tr>
<tr>
<td>Abstract (few lines):</td>
<td>This deliverable describes the orientation of the work package and presents the blueprint of the corresponding layer.</td>
</tr>
</tbody>
</table>
## Document History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Contributor(s)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3</td>
<td>15/03/2016</td>
<td>UAEGEAN</td>
<td>Preliminary draft</td>
</tr>
<tr>
<td>0.5</td>
<td>30/03/2016</td>
<td>UEAEGEAN, ICCS, JSI, UoW, NISSA</td>
<td>First draft - Discussion of open questions</td>
</tr>
<tr>
<td>0.7</td>
<td>20/04/2016</td>
<td>UEAEGEAN, AIT, TIS, KAPSCH</td>
<td>Intermediate version – for consortium review</td>
</tr>
<tr>
<td>0.9</td>
<td>27/04/2016</td>
<td>INTRA</td>
<td>Final Approval, QA</td>
</tr>
<tr>
<td>1.0</td>
<td>30/04/2016</td>
<td>INTRA</td>
<td>Final document for submission to the EC</td>
</tr>
</tbody>
</table>
Table of Contents

Definitions, Acronyms and Abbreviations ................................................................................. 7
Executive Summary .................................................................................................................... 8

1  Orientation and Forecasting Blueprint .................................................................................. 10
2  Traffic Forecasting Engine .................................................................................................... 12
   2.1  Sub-components of the forecasting engine ..................................................................... 13
       2.1.1  Simulator Engine .................................................................................................. 13
       2.1.2  Forecasting Models ............................................................................................... 13
       2.1.3  Social Miner for Traffic Predictions ....................................................................... 16
       2.1.4  Prediction Integrator ............................................................................................. 23
       2.1.5  Prediction Selector and Evaluator .......................................................................... 23
       2.1.6  Transportation Network status classifier ............................................................... 23
   2.2  Interaction with other Components ................................................................................ 24

3  Behavioural Econometric Models – Classification of Users .................................................. 25
   3.1  Introduction and literature review of Behavioural Models .............................................. 25
   3.2  Dynamic Travel Demand ............................................................................................... 28
   3.3  Departure Time Choice ................................................................................................ 29
   3.4  Big Data use for the development of behavioural models ............................................. 32
   3.5  The role of social network interactions on travel choices .............................................. 33
       3.5.1  The impact of the millennials and future trends ...................................................... 35
       3.5.2  Social Network Interactions on Travel Information Provision ................................. 38
   3.6  Inputs of the Model ........................................................................................................ 42
   3.7  Process of the Model ..................................................................................................... 43
   3.8  Outputs of the Model ..................................................................................................... 44

4  Complex Events Processing/Analysis ..................................................................................... 45
   4.1  The need for Complex Events Processing/Analysis ....................................................... 45
   4.2  High-level view on Complex Events Processing ........................................................... 47
   4.3  Benefits ........................................................................................................................ 50

5  Real time analytics .................................................................................................................. 52
5.1 QMiner ........................................................................................................................................ 52
  5.1.1 Mobility patterns detection service...................................................................................... 54
  5.1.2 Current state ...................................................................................................................... 57
  5.1.3 Next steps .......................................................................................................................... 61

6 References .................................................................................................................................... 62
# Table of Figures

**Figure 1:** OVERALL BLUEPRINT .......................................................................................................................... 11  
**Figure 2:** ARCHITECTURE OF TRAFFIC FORECASTING ENGINE ............................................................................ 12  
**Figure 3:** TWITTER GENERIC FRAMEWORK ........................................................................................................ 18  
**Figure 4:** PARTIAL VIEW OF THE SUMMARY TREE FOR THE STOCK MARKET APPLICATION (ARIAS, ET AL., 2013) .................................................................................................................. 22  
**Figure 5:** SUB-COMPONENTS AND INTERACTIONS ................................................................................................. 23  
**Figure 6:** INTEGRATED LV AND CHOICE MODEL FRAMEWORK, SOURCE: (BEN-AKIVA, ET AL., 2002) .................. 26  
**Figure 7:** MODELLING FRAMEWORK FOR DECISION MAKERS AND SOCIAL INTERACTION EFFECT ..................... 34  
**Figure 8:** CATEGORIES’ SHARE OF DESKTOP MINUTES 2015 EMEA (EUROPE, THE MIDDLE EAST AND AFRICA) .......................................................................................................................... 35  
**Figure 9:** SPOTLIGHT ON GLOBAL MILLENNIALS .................................................................................................... 36  
**Figure 10:** RIDE SERVICE APPS-UNIQUE VISITOR TREND IN THE U.S. .................................................................. 37  
**Figure 11:** CYCLING AND WALKING TRIPS - THE EFFECT OF MILLENNIALS ON ACTIVE TRANSPORT SHARE (SOURCE: BUEHLER AND PUChER, 2012) ................................................................................. 38  
**Figure 12:** SOCIAL PSYCHOLOGY AND ITS POTENTIAL INFLUENCE ON TRIP DECISIONS ........................................ 40  
**Figure 13:** MULTIMODAL BEHAVIORAL MODEL FRAMEWORK .............................................................................. 44  
**Figure 14:** COMPLEXITY OF DATA USED IN THE REROUTING PROCESS ................................................................. 45  
**Figure 15:** COMPLEXITY OF THE PROCESSING DATA IN THE TRANSPORTATION DOMAIN AND THE ROLE OF CEP ................................................................................................................. 47  
**Figure 16:** THE INTERACTION IN THE SYSTEM (GENERAL CASE) ............................................................................ 49  
**Figure 17:** QMINER. ..................................................................................................................................................... 53  
**Figure 18:** STEPS TOWARDS PERSONALIZED MOBILITY PATTERNS ........................................................................ 55  
**Figure 19:** DETECTED FREQUENT LOCATION ........................................................................................................ 57
List of Tables

TABLE 1: COMPARISON OF RELATED WORK ON CORRELATION AND PREDICTION USING SOCIAL MEDIA DATA .................................................. 16
TABLE 2: COMPARISON OF RELATED WORK ON TRAFFIC PREDICTIONS USING SOCIAL MEDIA DATA ................................................. 17
TABLE 3: MODELLING METHODS FOR PANEL DATA ..................................................................................................................................... 28
TABLE 4: TOP 5 STRENGTHS BY GENERATION ........................................................................................................................................ 36
TABLE 5: EMERGING CONSUMER DATA SOURCES (USA) .................................................................................................................................. 39
TABLE 6: ALGORITHMS .................................................................................................................................................................................. 58
TABLE 7: ALGORITHM RESULT OVERVIEW .............................................................................................................................................. 60
### Definitions, Acronyms and Abbreviations

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>API</td>
<td>Application Program Interface</td>
</tr>
<tr>
<td>ARN</td>
<td>Artificial Neural Networks</td>
</tr>
<tr>
<td>CEP</td>
<td>Complex Events Processing</td>
</tr>
<tr>
<td>GA</td>
<td>Genetic Algorithm</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphic User Interface</td>
</tr>
<tr>
<td>HCM</td>
<td>Hybrid Choice Model</td>
</tr>
<tr>
<td>IOT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>ITS</td>
<td>Intelligent Transportation Systems</td>
</tr>
<tr>
<td>LOS</td>
<td>Level of Service</td>
</tr>
<tr>
<td>MNL</td>
<td>Multinomial Logit Model</td>
</tr>
<tr>
<td>MST</td>
<td>Multivariate Structural Time-Series</td>
</tr>
<tr>
<td>OSN</td>
<td>Online Social Media</td>
</tr>
<tr>
<td>SUTSE</td>
<td>Seemingly Unrelated Time-Series Equation</td>
</tr>
</tbody>
</table>
Executive Summary

This deliverable describes the orientation of the work package and presents the blueprint of the corresponding layer. Initially, the flowchart and a short description are being presented. Each following chapter presents a certain module of the work package. First, the traffic forecasting engine is described, then the methodology of user classification is discussed through the theory of behavioral models, an introduction to the processing of complex events is then presented and the deliverable concludes with the real time analytics.

Chapter 2 presents the traffic forecasting engine. The engine provides predictions about the traffic flow and travel times in a short or medium term basis. The engine is based on several modules which are presented in this chapter along with the engine architecture. A literature review of best practices and state-of-the-art follows, describing parametric and non-parametric methodologies, as well as, hybrid systems. Then the chapter discusses the interaction with social media data and the relevant needed work in order to process social media (especially twitter) data into a useful form to the engine. The chapter concludes with a description of the form of the predictions and the relevant engine module.

Chapter 3 discusses the behavioral models that will be used in order to classify the users. Initially it describes the basic theory behind choice models and more advanced methodologies such as latent variable models. It then proceeds to describe more specialized forms of these models such as dynamic models and departure time models which are both relevant and useful to the deliverable. Social media interference with the behavioral models is also discussed. The chapter concludes with a brief description of the inputs, the processes and the outputs of the model.

Chapter 4 presents the handling and processing of complex events, which is useful to the deliverable and the project in general in order to reduce uncertainty in everyday travel. An investigation into the need for complex events processing and analysis is provided in the beginning sections of the chapter. The high-level view of this analysis is then presented, describing all the challenges of real-time processing and the architecture of communication between the user device and the server/processor. The chapter concludes by listing the benefits of the complex events analysis approach that is presented and provides examples of KPIs that can be optimized through this method.

Chapter 5 is concerned with the real-time analytics component of the platform. It introduces QMiner, an open source, big data, real-time handler of data that will be used in this work package. It proceeds to describe the mobility pattern detection service that will be developed and used in the project. An analysis of the internal components of the service is provided, with more specific information given on the location prediction service, the mobile app that
accompanies the service and the algorithms used. Evaluation and next steps conclude the section.
1 Orientation and Forecasting Blueprint

Figure 1 presents the overall blueprint of the work package. The middle part of the blueprint represents real user actions that take part as the user starts and continues to commute and conduct activities through his/her day. The upper layer depicts the inputs of data that can be briefly summarized as: Network data/Geodata, Sociodemographics, data from social networks, User behavior pattern data, Data from persuasive techniques and data from Traffic forecasting engine, econometric model and the routing engine. The clear connections with other work packages are: a) connection with WP2 for data input and b) connection to WP5 with data exchange for the persuasive techniques. In the middle part of the blueprint the user receives recommendations from the application for the specific destination he/she needs to travel to. The recommendations are for mode and for route and the user is allowed to select his/her preferred route. The choice is logged by the application and updates the model, the persuasive techniques and corrects simulation parameters through the updating/evolution/machine learning process. If an event happens en route, the routing engine pushes new information to the user. The user may request new information even if an event is not logged. When the trip for the specific OD is concluded, the routing engine is updated and the process initiates again for the upcoming trip.
Figure 1: Overall blueprint
2 Traffic Forecasting Engine

The Traffic Forecasting Engine provides short and medium-term predictions of traffic data like travel time and flow, as well as travel conditions for urban and inter-urban trips. It consists of the following sub components:

- a simulator engine,
- forecasting models,
- social miner to support traffic predictions,
- a prediction integrator,
- a prediction selector and evaluator; and
- a Transportation Network Status classifier.

Figure 2, provides a diagram for the architecture of the Traffic Forecasting Engine. The following subsections provide more details for each of the above components.
2.1 Sub-components of the forecasting engine

2.1.1 Simulator Engine
The Aimsun\textsuperscript{1} tool will be used as the simulator engine. The input required for this component is the transportation network, sections, nodes, control plans, traffic sites or OD pairs and traffic demand e.g. from sensors. The output will be predictions for traffic measurements as flow, speed and travel time. The extend and area coverage of the simulation models will depend on access to existing network models and data availability. Aimsun allows importing of networks from other modelling/simulation tools (e.g. Vissim, Visum, Synchro, Paramics) and this functionality will be employed where possible. Furthermore, Aimsun provides different programming interfaces (APIs) that allow automation of modelling tasks, as well as, dynamic control of the different elements of the simulation (vehicles, traffic demands, signal timings, etc.) at each simulation steps. These interfaces will be used for the integration of the simulator engine with the other sub-components of the forecasting engine.

2.1.2 Forecasting Models
Short-term traffic forecasting has been an integral part of most Intelligent Transportation Systems (ITS) and related research. Most of the interest has focused on developing methodologies that can be used to model traffic characteristics such as volume, density and speed, or travel times, and produce anticipated traffic conditions. In the beginning, almost all of the research employed ‘classical’ statistical approaches based on time series analysis for predicting traffic at a single point. However, these approaches have shown to be ‘weak’ or inadequate under unstable traffic conditions, complex road settings, as well as when faced with extensive datasets with both structured and unstructured data. Later, the research focused on empirical computational intelligence-based approaches, including Neural and Bayesian Networks, Fuzzy and Evolutionary techniques (Vlahogianni et al., 2014). Currently the existing methods belong to three broad categories: parametric, nonparametric and hybrid ones (we provide a literature overview in the next section).

This project uses methods that cover these three categories such as the ARIMA family, Kalman filtering, ANNs, and hybrid systems, and selects and evaluates the method that is appropriate for the specific application based on real-time traffic conditions. The basic input data for these models is traffic measurements, flow, speed, travel time but we also include other predictors (exogenous variables), e.g. weather data, incidents and features extracted from social media, e.g. tweets.

Different time aggregations will be tested (5 - 10 - 15 minutes) and various forecasting steps (one step ahead – multiple step ahead). Also, the predictions can be provided per single point, for a

\textsuperscript{1} https://www.aimsun.com
\textsuperscript{2} http://www.imdb.com/
\textsuperscript{3} https://dev.twitter.com/streaming/overview/request-parameters#filter_level
sensor, or per a path, route. The output will be predictions for traffic measurements. The following state-of-the-art models will be investigated and extended for the purposes of the OPTIMUM project.

2.1.2.1 Parametric Methodologies

In this category we have statistical methodologies that have been used in the field of time series forecasting. The Autoregressive Integrated Moving Average (ARIMA) model was used to predict short-term freeway traffic flow early at 70s (Ahmed and Cook 1979).

Many variants of ARIMA were proposed to improve prediction accuracy, such as Kohonen-ARIMA (KARIMA) (Vander Voort, Dougherty and Watson, 1996), subset ARIMA (Lee and Fambro, 1999), ARIMA with explanatory variables (ARIMAX) (Williams, 2001), vector autoregressive moving average (VARMA) and space–time ARIMA (Kamarianakis and Prastacos, 2003), and seasonal ARIMA (SARIMA) (B. M. Williams and L. A. Hoel, 2003) to better handle the seasonality in the data.

Another set of methodologies is related to kalman filtering and state space models. Stathopoulos and Karlaftis (2003), used Kalman Filtering (KF) for incorporating an extra dimension (spatial) as part of a time-series model. Flow input from spatially successive sensors formed a multivariate time series state space model with enhance performance compared to that of a univariate ARIMA model. Another important fact is that state space model is a multivariate approach, which makes KF suitable for network wide predictions (Whittaker, Garside, & Lindveld, 1997). Kamarianakis and Prastacos (2005), also enhanced time-series analysis by integrating spatial characteristics and data as part of an ARIMA model.

Regarding multivariate methodologies, Ghosh, Basu and Mahony (2009) proposed a multivariate structural time-series (MST) model using the seemingly unrelated time-series equation (SUTSE) to model the traffic flow time-series observations from multiple junctions within a congested urban transportation network.

2.1.2.2 Non Parametric methodologies

Parametric methods work best when the data is linear. Traffic flow exhibits stochastic and nonlinear characteristics and this fact made the application of nonparametric methods in the traffic flow forecasting field very appealing.

Davis and Nihan (1991) used the k-NN method for short-term freeway traffic forecasting and argued that the k-NN method performed comparably with but not better than the linear time-series approach. Chang et al. (2012) presented a dynamic multi interval traffic volume prediction model based on the k-NN nonparametric regression. El Faouzi (1996) developed a kernel smoother for the autoregression function to do short-term traffic flow prediction, in which functional estimation techniques were applied. Sun et al. (2003) used a local linear regression
model for short-term traffic forecasting. A Bayesian network approach was proposed for traffic flow forecasting in (Sun et al., 2006). An online learning weighted support vector regression (SVR) was presented in (Jeong et al., 2013) for short-term traffic flow predictions.

The majority of work in non-parametric methods is related to Artificial Neural Networks (ANN). Faghri and Hua (1995) and Lingras and Adamo (1996) developed models for predicting Annual Average Daily Traffic (AADT). Dougherty and Cobbett (1997) and Dia (2001) proposed various structures of ANNs for the short-term prediction of different traffic parameters, such as flow, speed and occupancy. Li and Rose (2011), developed an ANN for the prediction of mean travel times on motorways. The State Space Neural Network (SSNN) is considered as a variant of Elman Neural Network and has been applied to predict urban travel time (Van Lint et al., 2002, 2005; Van Lint, 2004; Stathopoulos and Karlaftis, 2003; Liu et al., 2006). The result demonstrates that the SSNN is superior to other prevailing algorithms in terms of accuracy (Liu et al., 2006). Previous research has identified that the time delay neural networks (TDNN) can achieve a higher travel time prediction accuracy compared with the SSNN (Shen et al., 2008). Long short-term Neural Network (LSTM) were used to predict travel speed and outperformed the other recurrent neural networks (SSNN, TDNN) (Ma et al., 2015).

2.1.2.3 Hybrid systems

There are many researchers that combined ANNs with other methodologies like optimization and Fuzzy Logic to improve the performance of ANNs.

Vlahogianni, Karlaftis and Goliás (2005) and Khosravi, Mazloumi, Nahavandi, Creighton and Van Lint (2011) used a Genetic Algorithm (GA) for optimising the structure and hyperparameters of an ANN for predicting flows and travel times in urban networks. He, Lu and Wang (2013) used Chaos Optimisation instead of SA to optimise the performance of a GA supported ANN. Zhu and Zhang (2009), constructed a layered forecasting algorithm using ANN and Self-Organizing Maps (SOM), while Chen, Grant-Muller, Mussone and Montgomery (2001) adopted SOM to classify traffic data as input to an ANN. In the former study, the layered structure was used to effectively support the forecasting task in situations where external events (e.g. incidents, social events, etc.) affected greatly recurrent traffic patterns.

Yin (2002) developed a fuzzy neural model which classified input data using fuzzy rules and mapped input to output data using an ANN, while Stathopoulos, Dimitriou, Tsekeris and Stathopoulos (2008) designed a fuzzy rule base system that combined ANN and Kalman filtering as part of a short-term forecasting model. Lu, Sun, Qu & Wang (2015) devised a deep-learning-based coefficients optimization algorithm based on fuzzy ANN and integrated as part of the forecasting model spatial and temporal correlations. Tan et al. (2015) proposed an aggregation approach for traffic flow prediction based on the moving average (MA), exponential smoothing.
(ES), ARIMA, and neural network (NN) models. The MA, ES, and ARIMA models were used to obtain three relevant time that were the basis of the NN in the aggregation stage.

### 2.1.3 Social Miner for Traffic Predictions

The social information miner sub-component will allow the exploitation of ITS related information embedded in social media textual sources, such as Twitter posts. This will be achieved via the text and meta-data analysis of the content streamed through the social media APIs. Several Natural Language Processing (NLP) techniques will be utilized to extract named entities (e.g. events, places), ITS related concepts (e.g. traffic congestion, delays) as well as emotional states (e.g. polarity, subjectivity) of the posted content. Social media meta-data (e.g., posts language, geo data, user profile meta-data) will be incorporated to complement this information, whenever they are available, and volume and temporal metrics (e.g., number of posts, time-series analysis) will be feature engineered to feed the forecasting engine.

#### 2.1.3.1 Related Work to Predictive Analytics Support with Social Media Data

The dramatic rise of social networking platforms such as Twitter or Facebook has resulted in the creation of huge repositories of user-generated content to a large extent available through the Internet (Arias, et.al., 2013). The wealth of social networking data has been incorporated during the last few years in a timely manner for identifying correlations between the additional social-based features and the domain-specific target features, and for the enrichment of predictive models. Table I depicts a comparison of previous works, where social media data has been exploited in several domains. Few examples include stock market forecasting (Bollen, et.al., 2011), (Zhang, et.al., 2010), box office predictions (Mishne, Glance, 2006), (Asur, Huberman, 2010), predicting elections (Tumasjan, et.al., 2010), and epidemics detection (Lampos, et.al., 2010).

Table 1: Comparison of Related Work on Correlation and Prediction Using Social Media Data

<table>
<thead>
<tr>
<th>Reference</th>
<th>Event</th>
<th>Predictive Model(s)</th>
<th>Corpus</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zhang, et.al., 2010</td>
<td>Several stock markets</td>
<td>Correlation Analysis</td>
<td>English tweets, gazetteer of mood tokens (e.g. hope, fear, worry)</td>
<td>Identified correlations between emotion tokens (hope, fear, worry) and DJIA stock index directions</td>
</tr>
<tr>
<td>Bollen, et.al., 2011</td>
<td>Dow Jones Industrial Average (DJIA)</td>
<td>Self-Organizing Fuzzy Neural Networks</td>
<td>10M tweets, stock market prices</td>
<td>Improved DJIA stock predictions with public calmness index of tweets</td>
</tr>
<tr>
<td>Mishne, Glance, 2006</td>
<td>Box office movie sales</td>
<td>Correlation Analysis</td>
<td>Blog posts with links to IMDB2, box office sales</td>
<td>Identified correlation between sentiment of blog post references to movies and their financial success</td>
</tr>
</tbody>
</table>

---

Table 2: Comparison of Related Work on Traffic Predictions Using Social Media Data

<table>
<thead>
<tr>
<th>Reference</th>
<th>Event</th>
<th>Predictive Model(s)</th>
<th>Corpus</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lin, et.al., 2014</td>
<td>freeway traffic speed in Buffalo-Niagara metropolitan area</td>
<td>Linear regression models</td>
<td>Weather, Twitter data, and traffic information</td>
<td>Twitter data has a relatively high sensitivity for predicting inclement weather (i.e., snow) during daytime. Twitter-based weather variables can help improve the predictive accuracy of the forecasting models</td>
</tr>
<tr>
<td>Tejaswin, et.al., 2015</td>
<td>Traffic Management Dashboard system for traffic insights for Bengaluru, Mumbai, New Delhi</td>
<td>Random forest classifier, clustering</td>
<td>Several APIs (Twitter streaming, World Weather Online, Google Geocoding,)</td>
<td>Classified tweets to either insightful or irrelevant to improve the Traffic Management Dashboard system in terms of traffic insights and alerts provision</td>
</tr>
</tbody>
</table>

More recently, there have been a few emerging works on the exploitation of ITS related information embedded in user-generated content of social data resources to support predictive analytics in traffic domain. Table 2 presents a summary of more related works to the objective the social miner component in the Optimum project, where social media data has been exploited in the fields traffic and transportation to support predictive analytics. Few examples include freeway traffic speed predictions (Lin, et.al., 2014), traffic incident clustering and prediction (Tejaswin, et.al., 2015), traffic incident management (Steur, 2015), real-time traffic congestion identification (Gong, et.al., 2015), traffic flow predictions (Ni, et.al., 2014), (Pathania, Karlapalem, 2015), and vehicle arrival time predictions (Abidin, et.al., 2015).
<table>
<thead>
<tr>
<th>Reference</th>
<th>Topic Description</th>
<th>Methodology</th>
<th>Twitter data impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Steur, 2015</td>
<td>Traffic incident management / detection on Dutch highways</td>
<td>Correlation statistics</td>
<td>Twitter streaming API</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Twitter data was not found to have sufficient spatio-temporal richness for traffic incident management</td>
</tr>
<tr>
<td>Gong, et.al., 2015</td>
<td>Real-time Traffic Congestion detection in Australia cities</td>
<td>Spatial-Temporal Clustering</td>
<td>Road network data for Australia, geo-targeted tweets</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Spatio-temporal clusters of tweets on roads collaborated to congestion detection in the studies areas</td>
</tr>
<tr>
<td>Ni, et.al., 2014</td>
<td>Traffic flow predictions during sport events in the Oracle arena in Oakland, California</td>
<td>Autoregressive, NNs, SVMs, k-NN models</td>
<td>California freeway traffic data, SM tweets from Twitter Streaming API</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Model enrichment with Twitter-engineered features improved prediction accuracy of regression models</td>
</tr>
<tr>
<td>Pathania, Karlapalem, 2015</td>
<td>Agent-based architecture to detect the movement of general public</td>
<td>Agent-based models, influence-probability models</td>
<td>Singapore’s metro train network data, Social Event Broadcaster data</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Based on a city simulation framework, agent-based broadcasting of events extracted from tweets is claimed to be a source of human travel information useful for predicting traffic flow</td>
</tr>
<tr>
<td>Abidin, et.al., 2015</td>
<td>Bus arrival time predictions</td>
<td>Kalman filter (KF) models, semantic analysis on tweets</td>
<td>Twitter streaming API</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Twitter-engineered features extracted by semantic analysis of tweets and fed as new input for route calculations and updates during KF model real-time data processing were found to be able to overcome data processing limitations in existing baseline models and improved accuracy of bus arrival time predictions</td>
</tr>
</tbody>
</table>

**Generic Framework**

- **Twitter data streaming and processing**
- **Exploratory Data Analysis**
- **Predictive Analytics**
- **Model Summarization**

*Figure 3: Twitter generic framework*
### 2.1.3.2 Twitter data streaming

Traditionally, text-based information retrieval research has focused on corpora that—though diverse in the type of documents (such as news articles, Web pages, or patents)—have a number of commonalities:

1. Each document consists of at least a few hundred words,
2. The content is (mostly) correctly spelled and grammatically sound, and
3. The information contained in each document is expressed in multiple ways.

In the case of microblogs such as Twitter, the need for an effective search filtration mechanism to find and filter relevant messages, or tweets, is evident, as Twitter receives more than a billion queries a day (Twitter Engineering Blog, 2011). Considering this from an information retrieval (IR) perspective, each Twitter message can be viewed as a document and for each user query, all messages that were created up to that point in time can then be considered for retrieval.

The Twitter Streaming API enables streaming real-time Twitter posts through a number of request parameters\(^3\). Two of these are ‘follow’ and ‘track’. The follow parameter offers the retrieval of Twitter messages that are posted by a pre-determined list of Twitter accounts, and can include a comma-separated list of user IDs, indicating the users whose Tweets should be delivered on the stream. The track parameter offers the retrieval of Twitter messages that contain list of phrases which will be used to determine what Tweets will be delivered on the stream. A phrase may be one or more terms separated by spaces, and a phrase will match if all of the terms in the phrase are present in the Tweet, regardless of order and ignoring case. By this model, commas can be thought of logical ORs, while spaces are equivalent to logical ANDs (e.g. ‘the twitter’ is the **AND** twitter, and ‘the,twitter’ is the **OR** twitter).

For the posting style on Twitter, users rely on different methods to adhere to the 140-character limit while conveying their intended meaning and emotions:

- Users tend to use abbreviations, remove vowels, drop articles, or use acronyms (Gouws, Metzler, Cai, & Hovy, 2011).
- Interjections and long sequences of repeated letters are widely used to express emotions.
- Hashtags are used both to engage in a discussion on a particular topic and to promote oneself (Laniado & Mika, 2010).
- Mentions are used when explicitly notifying others and when replying to others.
- URL shortening services are used to save characters when users include links. Such shortening removes valuable information such as the URL’s domain name (often used as quality indicator).

---

\(^3\)https://dev.twitter.com/streaming/overview/request-parameters#filter_level
In addition to the follow and track filtration mechanisms, the Twitter streaming API offers the retrieval of geo-located Tweets falling within the requested bounding boxes of longitude and latitude boundaries. A major challenge that needs to be dealt with when using Twitter data as a source of geographic information is the scarcity of tweets that are geo-referenced. Approximately 1% of all tweets are explicitly geotagged (Schulz et al., 2013). In an extensive literature review, Schulz et al. (2013) summarized twenty studies that dealt with this challenge of geo-locating tweets or Twitter users. In these studies, different spatial indicators were used in order to geo-locate a tweet or Twitter user. Most of the time, the tweet’s text was used by applying natural language processing (NLP) techniques on the terms in the text. An alternative for using natural language processing, that was used in the remainder of the studies is matching the terms in a tweet with a database of geographic locations, using a Gazetteer. An advantage of this approach is that it does not require training data and is much simpler (Schulz et al., 2013).

Instead of using the message text, some studies focus on using the location information that is sent with a tweet. However, an extensive study by Hecht et al. (2011) on the location field in tweets showed that the location field is not a very good spatial indicator on its own. Schulz et al. (2013) are the first that used a multi-indicator approach for geo-localizing tweets and Twitter users. They designed this multi-indicator approach because this method should be less vulnerable to missing or incomplete data. The work of Schulz and his colleagues is one of the best on geo-localizing tweets in recent literature. They managed to geotag 92% of all tweets with an average distance error of less than 30 kilometres. Hence, it seems that geo-tagging tweets without a GPS-coordinate is a very challenging job. In the scope of this thesis, tweets should be geo-tagged on a very small scale in order to be suitable for usage in incident management.

2.1.3.3 Twitter text processing

Pre-processing text data is the process of cleaning and preparing the text for classification (Haddi et al., 2013). Online texts contain usually lots of noise and uninformative parts such as HTML tags, scripts and advertisements. In addition, on words level, many words in the text do not have an impact on the general orientation of it.

Keeping those words makes the dimensionality of the problem high and hence the classification more difficult since each word in the text is treated as one dimension. Here is the hypothesis of having the data properly pre-processed: to reduce the noise in the text should help improve the performance of the classifier and speed up the classification process, thus aiding in the exploratory data analysis and predictive analytics phases.

The whole process involves several steps: online text cleaning, white space removal, expanding abbreviation, tokenization, stemming, stop words removal, negation handling and finally feature selection. All of the steps but the last are called transformations, while the last step applying some functions to select the required patterns is called filtering (Meyer et al., 2008).
For feature selection, there are several ways to assess the importance of each feature by attaching a certain weight in the text. The most popular ones are: feature frequency (FF), Term Frequency Inverse Document Frequency (TF-IDF), and feature presence (FP). FF is the number of occurrences in the document. TF-IDF is given by:

\[
TF - IDF = FF \cdot \log(N/DF)
\]

where \( N \) indicates the number of documents, and \( DF \) is the number of documents that contains this feature [15]. FP takes the value 0 or 1 based on the feature absent or presence in the document.

### 2.1.3.4 Exploratory Data Analysis

Modelling time series involves to a certain extent subjective judgement; nonetheless one can draw some general guidelines through statistical testing. In order to have some certainty of the model adequacy for Twitter as part of a traffic forecasting model, widely accepted tests can be run for exploratory data analysis to identify statistically significant relationships between the target time series data sets and Twitter (Arias, et al., 2013). This model adequacy evaluation can be in the form of, first, testing for nonlinear relationship among the targeted traffic data-based time series, and the time series built from Twitter-engineered features, which can either be simple aggregations of tweets (i.e., volume of tweets mentioning traffic event concepts), or more complex classifications (i.e., sentiment polarities and scores of tweets); and second, testing for causality from and to the Twitter series and the target time series, at different lags. Neglected nonlinearity is a multivariate test of nonlinearity to ascertain if two time series are nonlinearly related. This can be achieved with the neural network test for neglected nonlinearity developed by White (1989) and Lee et al. (1993). The basic idea is to perform a test of the hypothesis which states that a given neural network defines a perfect mapping between its input and output, and that all the errors are due to randomness. For tests of linearity, Teräsvirta linearity test (Teräsvirta, et al., 1993), which is based on White’s neural network test for neglected nonlinearity, can be used. The algorithm implementation is available in the tseries\(^4\) R library.

### 2.1.3.5 Model Summarization

In order to cope with the thousands of results obtained under the different experimental settings, there have been novel works which contributed to the development of a decision-tree-based summarization method of this information which we call summary tree (Arias, et al., 2013).

\(^4\) https://cran.r-project.org/web/packages/tseries/index.html
A summary tree identifies those parameter settings under which the forecasting prediction with Twitter is superior to the forecasting prediction without Twitter. In Arias et al. (2013), all possible combinations of a large number of experimental parameters have been grouped into three categories reflecting the main components of their forecasting experiments. Figure 4 depicts a partial view of a pruned summary tree, which shows those paths that lead to a clear win or loss in terms of number of experiments improving versus not improving the forecasting in the stock market application.

Figure 4: Partial view of the summary tree for the stock market application (Arias, et.al., 2013)

Sub-components and interactions

The social miner component for traffic predictions are composed of the following sub-components:

- Streaming data adapter to collect the tweets from the Twitter Streaming API using a geo-location filter
- Text processing engine to cleanse and pre-process the raw tweets.
- Relevance filtration classifier to identify the tweets that are about the traffic domain.
- Sentiment classifier to identify the polarity (positive, neutral, negative) of the pre-processed tweets.
- Aggregation engine to produce time series aggregations of traffic mentions and sentiment polarities from the pre-processed tweets.
- Feature engineering sub-component to derive extended features from the pre-processed tweets, such as extracting the named entities (people, organizations, locations).

As shown by Figure 5 below, the social miner component will mainly interact with the traffic prediction integrator sub-component and the central repository.
2.1.4 Prediction Integrator
This component will receive predictions from the other subcomponents of the Traffic Forecasting Engine, simulator engine and traffic models, along with features from the social miner for traffic prediction component. These predictions will be integrated and be transformed into common data formats, where applicable, for further analysis. Predictions will be performed for different traffic engineering parameters (e.g. flows, travel times, etc.), or (and) the status of the different parts of the network (e.g. highly congested, positive trend for flow, etc.).

2.1.5 Prediction Selector and Evaluator
Prediction Selector and Evaluator will get the predictions from the Prediction Integrator component and using statistical/machine learning techniques will give as output a prioritized list of predictions and a final (fused) prediction. Model summarisation techniques presented above will also be investigated for the development of this sub-component. The output will be a data set with the final predictions based on performance criteria and evaluations of the different models.

2.1.6 Transportation Network status classifier
Level of service (LOS) is a qualitative measure used to relate the quality of traffic service. LOS is used to analyse highways by categorizing traffic flow and assigning quality levels of traffic based on performance measure like speed, density etc. In general, the principle is to take the volume of traffic in one hour and divide by the appropriate capacity of the road type to get a volume/capacity (v/c) rating. There are 6 levels (A to F), where A and B may indicate a free
movement of the traffic (usually until a v/c of 35%-50%) and F over capacity (more than 100%). This component will try to correspond level of services with the appropriate model prediction output.

2.2 Interaction with other Components

Traffic Forecasting Engine will collect data from Optimum’s repository, and provide prediction traffic measurements to Traveller Behaviour component, Multimodal Routing Engine and Dynamic Charging component.
3 Behavioural Econometric Models – Classification of Users

3.1 Introduction and literature review of Behavioural Models

Discrete choice models have been used in the literature for decades, and are used to predict choices between two or more discrete alternatives, such as the mode of transport of a certain product. The main difference with models measuring continuous variables, such as number of trips, is that discrete choice models are concerned with the choice (the “which one”) as opposed to the quantity (the “how much”). Ben-Akiva and Lerman (Ben-Akiva and Lerman, 1985) describe the basic problem confronted by discrete choice analysis as: “the modelling from a choice set mutually exclusive and collectively exhaustive alternatives”.

Discrete choice analysis uses utility, dependent on observable independent variables and unknown parameters, estimating their values from actual choices made by decision makers. Discrete choice models attempt to model human behaviour, so they use the concept of random utility, because it is impossible to estimate a perfect model that will predict correctly all of human choices. So the utility of the alternatives are random variables and the output of the model, the probability that a certain individual will pick a certain choice is dependent on which alternative has the greatest utility.

A gap into the existing methodology was pointed out by Ben-Akiva et.al., in 2002. The traditional Random Utility Choice model, assumes that the model receives the input (individual characteristics, alternative attributes) and implicitly links the input data to the output (actual predicted choice). For this reason, attitudes and perceptions are not explicitly modelled. This means that while their general effect on the choice may be incorporated in the restricted model framework, the actual effect they have on the choice or the utility is not modelled and thus unknown to the researcher (Ben-Akiva, et.al., 2002). For this reason, a new framework incorporating the effect of attitudes and perceptions has been developed. In this framework ellipses represent latent constructs and the rectangles represent measurable variables (Figure 6).
Psychometric variables have been used by researchers (Wiley, 1973) (Joreskog, 1973) to model underlying constructs, latent variables that affect the human behaviour. This early research split the modelling concept of latent variables to a structural model (formation of the latent variables using observed variables) and to a measurement model (effects of the psychometric indicators to the latent variable).

Equations used in the integrated latent variable and choice models:

Suppose that we use Environmental Consciousness use of cars as a latent variable – Env

**Structural Model:**

\[
\text{Env}_n = X_n \gamma + \omega_n \quad \omega \sim N(0, \Sigma \omega)
\]

\[
U_n = X_n \beta + \text{Env}_n \delta + \varepsilon_n \quad \varepsilon \sim N(0, \Sigma \varepsilon)
\]

Where:

- \( \text{Env}_n \): Latent Variable
- \( X_n \): Explanatory Variables
$\omega_n, \varepsilon_n$: Vectors of random disturbance terms

$\beta, \gamma, \delta$: Unknown parameters

$U_n$ is the utility of the alternative, $\beta$ is the vector of the observed variables and $\gamma$ is a diagonal matrix of unknown parameters associated with the latent variable $Env_n$, $\omega_n$ is a vector of random disturbance terms associated with the utility terms. The choice model is assumed to be based on utility maximization:

$$y_i = 1, \text{ if } U_i = \max\{U_i\}, \quad i = \text{all alternatives}$$

$$y_i = 0, \text{ otherwise}$$

$y_i$: the choice indicator

**Measurement Model:**

$$I_{Env_n} = \alpha + \lambda Env_n + \nu_n \quad \nu \sim N(0, \Sigma_\nu)$$

Where $I_{Env_n}$ is a vector of attitudes and perceptions, $\alpha$ a vector of unknown parameters that indicate the association between the responses to the scale, $\lambda$ are vectors of unknown parameters that relate the random variable to the indicators, $\nu_n$ is a vector of random error terms.

**Likelihood function:**

$$f(\nu_n, I_{Env_n} | X_n; \alpha, \beta, \gamma, \delta, \lambda)$$

$$= \int_{Env_n} P(\nu_n | Env_n; \delta, \Sigma_\epsilon) f(I_{Env_n} | Env_n; \lambda, \Sigma_\nu) f(Env_n | X_n; \gamma, \Sigma_\omega) dEnv_n$$
Latent variables can classify users to discrete classes depending on their attitudes and perceptions. This has been done extensively in transportation literature, with a lot of work focusing on classifying the users into classes indicated by attitudes and perceptions and describing or explaining the user behaviour. For example (Anable, 2005) used psychometric questions to create six discrete user profiles: Malcontent motorists, Complacent car addicts, Die hard drivers (completely devoted to driving), Aspiring environmentalists, Car-less crusaders and Reluctant drivers.

3.2 Dynamic Travel Demand

Very limited work directly applied to dynamic travel demand forecasting was found. However, the examples that exist were helpful in showing potential methodological approaches in this area. Given that panel data is the main potential data source for dynamic modelling of travel demand, the literature review place emphasis on this topic. Panel data is suitable data needed to develop dynamic relationships. In statistics and econometrics, the term panel data refers to multi-dimensional data frequently involving measurements over time. Table 3 provides a summary of different methods that can be used for panel data modelling and to what purposes they can be applied.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Dependent variable</th>
<th>Dynamics considered</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Aggregate time-series analysis</td>
<td>• Time elapsed</td>
<td>• Non-stationarity</td>
</tr>
<tr>
<td>• Markov chains</td>
<td>• Change or not</td>
<td>– Socio-economics</td>
</tr>
<tr>
<td>• Switching models</td>
<td>• Choice</td>
<td>– Transport system</td>
</tr>
<tr>
<td>• Choice models</td>
<td>– Continuous</td>
<td>– Life events</td>
</tr>
<tr>
<td>• Multi-level analysis</td>
<td>– Discrete (qualitative, e.g. mode)</td>
<td>– Time explicit effect</td>
</tr>
<tr>
<td>• Event history analysis (duration or survival models)</td>
<td>– Discrete (ordered, e.g. none, once a week, more than once a week)</td>
<td>– Lag (and lead) responses</td>
</tr>
<tr>
<td>• Structural equation modelling</td>
<td></td>
<td>– State dependence</td>
</tr>
<tr>
<td>–Path analysis</td>
<td></td>
<td>– Markovian</td>
</tr>
<tr>
<td>–Latent growth</td>
<td></td>
<td>– Occurrence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– Duration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– Serial correlation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>– Unobserved fixed preferences</td>
</tr>
</tbody>
</table>
On the collection of panel data, Ma, K. and Chatterjee, K. (2007) conducted an analysis of a panel survey data set obtained for the London congestion charging scheme. The survey obtained information on travel perceptions and behaviour six months before and after the scheme’s introduction. The analysis developed simple binomial models for whether car drivers changed mode or not and it was found that the limited information collected on the specifics of behaviour and on attributes of modal alternatives, as well as the limitations inherent in two wave panel data, prevented development of models with more sophisticated dynamic characteristics.

3.3 Departure Time Choice

An individual’s travel behaviour is characterised by a series of important decisions including, but not limited to, the choice of frequency, location, mode and route of travel. Most urban travel models explicitly address these four dimensions through the four-step model system: trip generation, distribution, mode choice and assignment while time-of-day issues receive less attention. Popuri, Ben-Akiva and Proussaloglou, in their work (2008) explore time of day decisions and their impact of shorter term congestion mitigation strategies such as road pricing. They also highlight the necessity of higher temporal resolution in travel demand forecasts in assessing traffic emissions and air quality. They have developed a methodology to address the above mentioned issues. They describe the data inputs required for modelling time of day decisions and they present the model estimation procedure and empirical results from Tel-Aviv case study. The main features of their modelling approach include using half-hour time intervals, accounting for schedule delay in the absence of desired arrival and departure time data, and modelling the 24-hour cycle.

If \( a \) and \( d \) represent the arrival and departure time slots respectively, the joint utility \( U(a,d) \) is modelled as:

\[
U(a,d) = \alpha(T_{a}^{\text{main}}) + \beta(T_{d}^{\text{main}}) + \gamma(T_{d}^{\text{main}} - T_{a}^{\text{main}}) + \tau_{1}^{\text{home}} (T_{a}^{\text{main}}) + \tau_{2}^{\text{home}} (T_{d}^{\text{main}}) + \delta_{1}^{\text{dir}} \ln(\Delta_{a}) + \delta_{2}^{\text{dir}} \ln(\Delta_{d}) + \epsilon_{a,d}
\]  

(1)

Where,

\( U(a,d) \) represents the utility of arriving at the main stop in time slot \( a \) and departing from the main stop in time slot \( d \), such that \( a, d \in \{1,2,...,36\} \) (36 time slots), and \( a \leq d \) (the arrival time at the main stop needs to precede the departure time from the main stop). Thus, are defined a
total of 666 (=36 * (36+1)/2) utilities corresponding to the various arrival and departure time combinations;

\[ T_{a\,main} \] and \[ T_{d\,main} \] represent the mid-points of the arrival and departure time slots, measured in hours from an arbitrary time such as midnight;

\[ a(T_{a\,main}) \] and \[ \beta(T_{d\,main}) \] represent the arrival and departure time functions that capture, among other things, the effect of schedule delay in the absence of data on desired time of arrival and departure.

\[ \gamma(T_{a\,main} - T_{d\,main}) \] represents the duration function for the primary work tour.

\[ t_{home\,|T_{a\,main}} \] is the travel time from home to the main stop for the arrival time \( T_{a\,main} \) at the main stop, and \( t_{home\,|T_{d\,main}} \) is the travel time from the main stop to home for the departure time \( T_{d\,main} \) from the main stop. \( \tau_1 \) and \( \tau_2 \) are multiplicative parameters that need to be estimated;

\( \ln(\Delta_a) \) and \( \ln(\Delta_d) \) are the log-size measures that are introduced to account for the unequal size of the 36 time slots that are used to discretise continuous time. \( \Delta_a \) and \( \Delta_d \) are defined as the number of half-hour intervals in the arrival and departure time slots respectively. \( \delta_i \) and \( \delta_2 \) are multiplicative parameters that need to be estimated. The size terms are essentially correction terms applied to the first and the last time slots, which are both larger than 30 minutes in duration.

Finally, \( \epsilon_{a,d} \) are the error components assumed to be independent and identically Gumbel distributed.
Also Xia Jin, Alan Horowitz in 2008 are concentrating their research on the temporal nature of trip making into the travel demand modelling process. While the vast majority of prior studies have focused on daily urban trips, they explore the timing/scheduling decision-making behaviour for long, occasional and exceptional travel, rather than habitual, repetitive trips. An intensive preference survey was conducted to help expose those salient factors that affect time-of-day (TOD) choice, and help understand the prioritisation among the variables and constraints. A multinomial logit model was then developed from the 2001 National US Household Travel Survey daily-trip survey data. Various trip activity, personal and household characteristics were examined. The major hypothesis tested in this study is that TOD choice for long distance trips is highly dependent upon the characteristics of the trip/activity, the circumstances of the traveller, and the level of service of the transportation system; and that it is possible to forecast the temporal distribution of demand with greater accuracy when these factors are incorporated into choice models. After the preference survey, a multinomial logit model (MNL) was developed to further test the hypothesis. The procedure forecasts the probability of one TOD being chosen for each long-distance trip based on a set of taste parameters and the attributes of the alternatives and the decision-maker. The model used in this study is described in the equation below:

\[
Pr(\text{TOD}) = f(T, A, P, H)
\]

where

TOD = TOD period choice for the long trip,
T = trip related factors such as purpose, mode, travel time, traveling companions,
A = activity related factors such as activity duration,
P = personal characteristics such as age, gender, education level,
H = household characteristics such as income, size, auto-ownership, presence of young child

The results of this study showed that the departure time choice for long trips was strongly affected by how long the trip would be, how much time would be spent at the destination, whether traveling on a weekday or weekend, whether traveling alone or with other companions, and whether traveling with young children.

To predict the dynamic travel demand is to model travellers’ departure time choice. This is often done by applying an exogenous response curve stating the percentage of departures in each time interval. Since some origins will be earlier under threat than others, such a response curve is typically predicted for each origin separately. The departure response curve has been assumed to follow many different distributions. Some examples are instantaneous departure (Lewis 2001;
Chen and Zhan 2004; Chiu et al. 2006), a Uniform distribution (Liu et al. 2006; Yuan et al. 2006), a Rayleigh distribution (Tweedie et al. 1986), a Poisson distribution (Cova and Johnson 2002), a Weibull distribution (Jonkman 2007; Lindell 2008) or sigmoid curve (Kalafatas and Peeta 2009; Xie et al. 2010). The Weibull distribution and sigmoid curve are most often used and claimed to be most realistic. The Weibull distribution is given by

\[ D(t) = 1 - \exp(-\beta t^\gamma) \]  

(2)

3.4 Big Data use for the development of behavioural models

Real time Big Data processing has become important for various applications related to mobile and location-based services. For example, location data can be gathered in different ways and with different precision, depending on the current context and business goals. For the transportation domain the efficient integration of heterogeneous real-time data in order to achieve complex situational awareness, by taking into account the privacy concerns is especially challenging.

There are two main streams of research which are related to the research in Big Data processing: Agent-based cloud computing and Internet of Things (IoT). Agent-based cloud computing is a paradigm that identifies several common problems and provides several benefits by the synergy between MAS and cloud computing. Software agents can be used as basic components for implementing intelligence in clouds, making them more adaptive, flexible, and autonomic in resource management, service provisioning and large-scale application executions [Tal11].

IoT semantically means a world-wide network of interconnected objects uniquely addressable that ensure the exchange and sharing of information in ITS. The basic idea of this concept is the pervasive presence around us of a variety of things or objects (radio frequency identification tags, sensors, actuators, mobile vehicles, etc.), which, through unique addressing schemes, are able to interact with each other and cooperate with their neighbours to reach common goals. IoT provides for ITS two main things: 1) its data acquisition function provides more comprehensive traffic data; 2) provides a good channel for traffic data transmission. Therefore, ITS based on IoT has broad prospects of development and expansion space [Xia11].

Classical methods of data processing and mining were centralised. In opposite, Big Data is constantly updated and collected in physically distributed storages [Fio11]. Therefore there is an inherent need to develop effective Big Data processing algorithms using decentralised architecture that takes into account space and time distribution of data.
We have identified a gap in a novel approach/architecture for real-time big data processing (based on Storm, [Liu14]) that senses ad-hoc dynamicity/changes in real-time data and if needed it changes the processing architecture. Such dynamicity in processing represents an important innovation, especially because some of the changes (in the processing architecture) can be learned automatically, enabling the system to continually improve its real-time processing capacity. The system will be based of extending past work of NISSA in the domain of CEP [Luk02]. The approach include efficient storage of the data, enabling easy access to data from different devices/platforms.

Another important big data technique to forecast the behaviour of travellers is the analysis of mobile phone location based data. Analysis of location based data is based on associating the characteristics of human movement with places and infrastructures locations. It is thus possible to predict these movements by seeking the relations between these displacements and other available information like the user’s profiles and the site of the infrastructures. Because of the complexity of the human displacements characteristics and in the absence of a reliable law of movements prediction, data mining can constitute a solution to the prediction.

The analysis of data from cellular phone systems in order to study long-distance travel patterns has been applied in Israel in 2009. The approach allowed passive data collection on many travellers over a long period of time at low costs. The method was specifically designed to capture long distance trips, as part of the development of a national demand model. The method allowed the construction of Origin-Destination tables directly from the cellular phone positions. It is not intended to replace household surveys, but to complement household survey data.

3.5 The role of social network interactions on travel choices

It is well-known that individuals’ choice behaviour is often influenced by the existence, opinions, choices and behaviours of other people (van den Bos et al. 2013; Rose and Hensher 2004; Brock and Durlauf 2001; Manski 1993) or generally by the social environment of the decision maker. In sociology and psychology, there is much empirical evidence confirming the effect of social interaction. Lately, the effect of social interaction and social influence on individuals’ decision-making has attracted attention in the transportation sector as well.

Polydoropoulou, Kamariani, Ben Akiva (2014) presented a general methodology and framework for including social interaction effect into HCM (Hybrid Choice Modelling). Based on the findings in psychology and neuroscience research that the individual’s decisions are indirectly influenced by the social environment, as it affects the individual’s psychological state (van den Bos et al. 2013; Homberg 2012), the developed method provides insights for modelling the effect of social interaction on the formation of psychological factors (latent variables) and on the decision-making process. Thus, the social environment is a latent variable that represents social
interaction with the decision maker and it is included as a component to the latent variable regarding the decision maker, which in turn is included directly in the choice model. The methodology that they proposed, requires the estimation of an integrated multi-equation model consisting of a discrete choice model, the latent variable model’s structural and measurement equations regarding the decision maker and the latent variable model’s structural and measurement equations regarding the social environment. The methodology is tested within the context of a household aiming to identify the social interaction effects between teenagers and their parents regarding walking-loving behaviour and then the effect of this on the mode to school choice behaviour. The methodology provides the ability to researchers to specify as many latent variables for the social environment as they want. For example, different latent variables could be used for parents, siblings, friends, colleagues etc., each one representing a different social network. Moreover, this could provide insights about which social network affects more the behaviour of the decision maker. In Figure 7 we present accordingly this modelling framework for decision makers and how there are affected by their social environment.

Figure 7: Modelling framework for decision makers and social interaction effect
An emerging issue today is the extended use of the internet, smartphone devices and social networking. The use of these technologies and applications introduce new ways in the flow of information that can affect also decision making regarding mobility and transport issues. The 2015 Digital Future in Focus Reports share key figures and trends in digital behaviour across the global markets measured by comScore. It shows that social media takes 22% of the desktop minutes spent by users in the EMEA region (see Figure 8).

![Figure 8: Categories’ share of desktop minutes 2015 EMEA (Europe, the Middle East and Africa)](image)

Source: comScore report: 2015 Global Digital Future in Focus

Susan Pike (2015) explores social networks and travel behaviour among students in two university settings. The main aims are to understand whether the role of social influence in mode choice differs with respect to the transportation environment. Based in her case study she concludes that the effects of social influence in transportation mode choice is limited; measured by an ego-network consisting of five people that the respondent has some kind of personal or professional relationship with. She also highlights that individuals who have more options may be more influenced by their social networks, while individuals with relatively little option for commute modes tend to have limited reactivity to the mode use of others in their social networks. Hence, there is a varying potential for social network influence in transportation mode choice.

### 3.5.1 The impact of the millennials and future trends

Millennials – also called as Generation Y – are today aged 18-24. They are one of the largest population segments nowadays, on pair with Baby Boomers. These young consumers are the largest segment of smartphone owners. The company Nielsen - that studies consumers and gives a view of trends and habits worldwide - shared in 2015 some data about U.S. smartphone market share by age. According to this data, over 85% of Generation Y owns Smartphones. In the U.S., 171.5 million people (71%) own such a smartphone device. Smartphones have become the
A staple of everyday life and the on-the-go tool of choice for consumers looking to catch up on emails, tap their social networks or even tweet about a recent sports game.

The global trend for the millennials is that they spend more time on internet through smartphones than personal computers, comparing to the average user. In Figure 9 we can see a graphic representation of this trend and also an analysis showing that millennials use their smartphone for browsing entertainment and social media webpages more than all the other categories of users.

![Figure 9: Spotlight on global millennials](source: comScore MMX &Mobile Metrix, 2015)

According to Gallup Inc., an American research-based and global performance-management consulting company, the top strength of the Millennials is adaptability. Table 4 we present the list of the Top Five strengths by Generation:

**Table 4: Top 5 strengths by generation**

<table>
<thead>
<tr>
<th>1</th>
<th>Adaptability</th>
<th>Input</th>
<th>Responsibility</th>
<th>Responsibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Input</td>
<td>Achiever</td>
<td>Achiever</td>
<td>Harmony</td>
</tr>
<tr>
<td>3</td>
<td>Responsibility</td>
<td>Responsibility</td>
<td>Adaptability</td>
<td>Empathy</td>
</tr>
<tr>
<td>4</td>
<td>Achiever</td>
<td>Learner</td>
<td>Developer</td>
<td>Consistency</td>
</tr>
<tr>
<td>5</td>
<td>Context</td>
<td>Relator</td>
<td>Empathy</td>
<td>Achiever</td>
</tr>
</tbody>
</table>

*source: Gallup Inc., 2014*
The fact that millennials are more adaptive to new situations and the technology, as part of their behaviour, could lead into major changes in way they will make their transport choices in the near future.

Accessing on-demand services has also become habitual for many people, and Uber & Lyft lead the way for on-demand transportation. Uber and Lyft have exploded in popularity over the past year by improving upon an existing offline behaviour (i.e. hailing a taxi) via the large online networks that they’ve built. Both apps benefit from users who have a higher tendency to be a Millennial and live in a single-person household.

![Figure 10: Ride Service Apps-Unique Visitor Trend in the U.S.](source: comScore Mobile Metrix, U.S., Age 18+)

Compared to EU countries, especially western European countries like the Netherlands or Denmark, these numbers seem low. The percentage of trips conducted by bicycle or walking in the US was less than 5% in 2008, while at the same time it was over 50% in the Netherlands and almost 35% in Denmark. The effect of the latest generation of people is prominent in these EU countries too. As Figure 11 depicts, in the mentioned western European countries the millennials tend to use more walking and cycling as means of transport. In fact, the effect of millennials using active modes is stronger than the US in those particular countries.
The millennials are leading the way to new mobility patterns and new ways of transport choices for the near future. The effect of the financial crisis, the advancement in the technologies of communication, the seamless access to information via smartphones, are some new facts that decision makers should take into account in transport planning.

3.5.2 Social Network Interactions on Travel Information Provision

The provision of travel information is becoming a commodity. There is a variety of modes for the provision of this kind of data, prescriptive, descriptive, on maps, by voice, etc. In Table 5 we present some providers of travel information data in the U.S. who are using the latest technology advancements.
### Table 5: Emerging consumer data sources (USA)

<table>
<thead>
<tr>
<th>Technology</th>
<th>Example Provider</th>
<th>Primary Market Focus</th>
<th>Potential Public-Sector Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-vehicle navigation device</td>
<td>TomTom</td>
<td>Navigation and real-time traffic information</td>
<td>Transportation system performance, repetitive travel patterns</td>
</tr>
<tr>
<td>In-vehicle service</td>
<td>OnStar</td>
<td>Location-based services</td>
<td>Origin–destination data, parking, transportation system performance</td>
</tr>
<tr>
<td>Mobile phone tower-to-tower</td>
<td>AirSage</td>
<td>Traffic data, population movement data</td>
<td>Origin–destination data, population movement, long-distance travel times</td>
</tr>
<tr>
<td>handoffs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smartphone application</td>
<td>INRIX</td>
<td>Real-time and predictive traffic information</td>
<td>Transportation system performance, origin–destination data, trip-making patterns</td>
</tr>
<tr>
<td></td>
<td>Traffic Google</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Maps</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

According to C.Bartle and her paper “Spreading the Word”, the use of “formal” travel information pertaining to costs, routes, journey times, or real-time transport disruptions, and its role in travel behaviour (for example, choice of mode, route or departure time) has been widely studied, but little is known about the part played by “informal” information, shared through word-of-mouth amongst friends, family, colleagues and other social networks, in relation to everyday travel. Furthermore, considerable investment has been made over recent decades in the development of sophisticated “advanced traveller information systems”, delivering formal, top-down information through media such as online journey planners, but less attention has been paid to parallel developments in the diffusion of bottom-up, user-generated information through „electronic word-of-mouth“ on the internet (acknowledged in the field of marketing as a growing source of influence on consumer behaviour). The role of word-of-mouth information diffusion within everyday travel behaviour and its emerging applications in the field of online traveller information, within a framework of social-psychological theories of behaviour and decision theory has been examined by C.Bartle. The exploration of social-psychological factors underlying the social transfer of traveller information led to an expansion of existing theory, whilst her research also generated practical recommendations for the wider incorporation of “social design features” into certain forms of traveller information system.

Her schematic diagram on social psychology and its potential influence on trip decisions is presented in Figure 12.
Recommendations for the incorporation of “social design features” into other forms of advanced traveller information system highlighted in particular the benefits of combining user comments with routes generated by automated planners. User-generated route information was considered to be especially relevant for cycling and walking, but general user comments on transport services were also considered to be potentially useful for public transport. Comments could be posted on the map alongside public transport routes, and it was thought that this might be useful not only for other travellers (especially those who were new to, or considering a mode of transport), but also in terms of providing feedback to transport operators. Similarly, user comments posted to the map by cyclists and pedestrians (as well as wheelchair and pushchair users) could communicate both good and bad features of the route to those responsible for maintaining them. The community-building aspects of this type of system were thought to be beneficial in terms of motivating and sustaining cycling, walking, lift-sharing and the use of public transport. Community-building was thought most likely to occur within distinct groups where members shared a sense of identity. In order to sustain enough interaction to create or maintain such a sense of community, approximately 50 contributing members were thought to be required. Four “types” of website user were identified in the Cycology project, and it was
concluded that “active enthusiasts” were particularly important in sustaining interaction, although “active sceptics” might also be expected to contribute for pro-social reasons, and “interested lurkers” provided a reservoir of potentially more active users.

According to P.Gault, C.Cottrill, G.Yeboah, J.Nelson and J.Anable (2015), the communication of information between public transport service providers and their customers has changed dramatically with the advent of GPS, smartphones, and digital social networking. Far from the days of reliance upon static timetable information posted at stops and paper-based maps, providers have widely embraced real-time updates, internet-based journey planners, and a variety of other digital methods to communicate relevant and timely information to their customers (Tang and Thakuriah, 2012). In perhaps no area is this transition more evident than in the increasing use of the microblogging site Twitter for purposes of customer communication. The Twitter brand has become widespread in the public transport field (Camacho, Foth, and Rakotonirainy, 2013), as it provides a cost-effective and reliable method of sharing information that may be time-sensitive or impact upon large sections of the transport network. Twitter’s status as a public forum, however, has also introduced a myriad of concerns to the public transport arena, and has created a need for public transport operators to re-evaluate and modify their operating procedures regarding communication. Equally critical is the need to ensure the accuracy and reliability of posted information, as the mechanics of Twitter increase the likelihood that Tweets will be shared and spread far beyond the original forum (Ye and Wu, 2010). In their paper P.Gault, C.Cottrill, G.Yeboah, J.Nelson and J.Anable (2015), they examine the use of Twitter in the context of a large event, specifically considering how Twitter was used to manage relevant information and broadcast a coordinated message regarding transport and travel disruption. The 2014 Commonwealth Games (hosted by Glasgow, Scotland) are used as a case study. They focus on the mechanics of using Twitter in such a coordinated manner, using a combination of interview data and analysis of Tweets. Findings include positive effects of enabling competing transport operators, who would normally work in separate organisations, to come together for a shared goal under pressured circumstances; as well as challenges associated with coordinating the social media message, the on-ground situation, and user responses in a public forum. The social media model used for transport information dissemination during the Games demonstrates what can be achieved when separate entities coordinate themselves more efficiently, and also provides indications of areas where additional planning and evaluation are needed. Whilst such a process may not always be necessary outwith such a large-scale event, it demonstrates the role a technology solution such as Twitter may play in helping to facilitate a dialogue between the public, transport operators and other information providers.

In third world countries sharing of traffic information through social networks has become an important game changer. Also in Latin America and more specifically in Venezuela, social networking is an important way of communication for the people and is being established as an
alternative means of information to the mainstream media. The analytical company comScore stated that the audience of Venezuela in Twitter boomed to 19% when the late president Hugo Chávez created an account there. Venezuela now has the 4th highest percentage of Twitter users. An interesting detail is that in Venezuela the social networks are also used for announcing dangerous road spots where criminal activity is taking place.

3.6 Inputs of the Model

In order for the Behavioural Econometric model to work certain input must be provided. The basic input of the model can be viewed in the Multimodal Behavioural Model Framework (Figure 3). Attributes of alternative modes, habitual travel patterns of individuals and socioeconomic data of individuals are the main input provided to the model. Additional input includes psychometric questions (1 to 7 agree/disagree Likert scale), which will act as the indicators for the latent variables of the model (left side of Figure 7). Also data such as mode availability, weather conditions, schedule flexibility will be useful in order to define the latent choice set and to better design the stated preference experiments.

Socio-demographic data will be gathered through the questionnaires of the project. Application will provide the travel patterns of the individuals and the local authorities will provide data about the levels of attributes of different modes. For example, in a certain area mean travel time may be 20 minutes so the range of travel time attribute in the experiment will be from 10 to 40 minutes. The three data components described will be the base for the variables used in the model.

Psychometric questions embedded in the questionnaire will be used to determine the classes that users belong to. Each class is indicated by a number of attitudinal questions and has its own socio-demographic statistics that reveals which users are more likely to belong to this class, constructing the user profile of the class. For example, user that answer that they agree or mostly agree with pro-environmental statements form the “eco-friendly” class. This class has certain socio-demographics, for example, young aged, highly educated, high salary individuals. Then the model may predict the possibility of a user belonging to this class independent of the existence of attitudinal questions for this user.

When considering individual cases, not all modes or combinations of modes are available to all users, every day. A lot of factors affect the availability of a choice. Weather conditions may hinder the possibility of an active transport choice or the non-existence of a frequent bus schedule on the study area may dissuade an individual from choosing to ride with bus. Data like this, schedule, weather conditions, density of public transport network, other constraints will be used to determine a specific latent choice set for every user. This means that each user will be
presented only with the choices she could actually choose from, actively penalizing the utility of an alternative based on its perceived availability. (Vij, et.al., 2013)

3.7 Process of the Model

The model simultaneously estimates three different components, as described in the previous section: a) the structural model – which is the formulation of the latent classes based on socio-demographic data b) the measurement model which is the assignment of content to these latent classes, based on the indicators and c) the discrete choice model – the part that models the actual choice of the individual.

The model contains two latent variables and a latent choice set. The first latent variable is environmental consciousness, measured by indicators such as: I feel conscious about climate change or I try to use the car less to reduce my CO$_2$ emissions. The second latent variable is time reliability. There is also a latent choice set present in the model process. The latent choice set is constructed by the perception of different mode availability for each user, location and weather conditions and it affects the content of the stated choice experiments presented to each user. The final utility for each mode and subsequently the probabilities of choosing each mode is affected by the latent variables and directly by the socio-demographics.
3.8 Outputs of the Model

1. Probabilities of choosing a mode or a combination of modes
2. Population segmentation – User profiles
3. Elasticity: The model can provide individual and aggregate elasticity. For example, an individual’s choice probability in the event of cost or travel time change or the reaction of a certain user profile to changes
4. Values of time
4 Complex Events Processing/Analysis

4.1 The need for Complex Events Processing/Analysis

Recent development in the transportation domain has shown an enormous potential of new sensor technologies and real-time processing for different types of process optimization. However, exploiting full potential of sensor technologies requires new approaches for the interpretation, integration and processing of sensor information, including transformation of the computing on the mobile as well as server/cloud part. In the case of rerouting process this is even more complicated since the data is very huge and heterogeneous. Following Figure 14 illustrates this situation.

There are two main challenges to be resolved:

C1) Need for closed-loop event processing: An important problem in transportation domain is the complexity of the situations that can happen and the need to be resolved immediately. Indeed, due to dynamicity of the changes in the environment, it is very important to react on the problems that might appear very promptly and efficiently. For example, a deviation of the temperature around the allowed threshold during the transportation can be critical for the quality of food and should be processed very fast. It means that some complex processing (e.g. trend detection) should be performed on the point of care (truck, locally) in order to support requested agility. On the other hand, a truck is involved in a complex logistic process and the
reaction should be synchronized with the global context. For example, if there is a problem with the health status of the (perishable) goods, the problem (or early indicators) should be communicated to the global context, that should (1) decide about the criticality of the situation, (2) find the solutions to resolve it and (3) communicate back to the local processing. Note that if needed this feedback loop can be repeated several times until the optimal solution found.

Therefore, the main issue is that the local processing/analytics should be related to some global planning based on the information which is collected centrally. It means that we need a kind of dynamic and distributed complex event processing that encompasses local and global event processing as a closed loop. Due to criticality of some transportation situation (e.g. traffic accident) both properties are of extreme importance:

- the dynamicity: it is required to react very promptly
- the distributed processing: combining local and global processing is needed in order to make the best decisions, in the case of some anomaly situations.

We can envision two types of situations that are of interest for this type of processing:

- anomaly situations that happen locally in the transportation (e.g. some problems with the truckload) and can be resolved properly only by consulting global context (e.g. to understand if such goods can be delivered at another place – in the proximity of the local position of the truck)
- situations that happen outside the transportation chain (e.g. a customer has an ad-hoc request for delivery/pick up) and can be resolved, by considering global context, through a local action, e.g. that a truck (driving in the proximity of that customer) should change the route and pick-up the goods from the customer.

C2) Need for the advanced data analytics, esp. detection of unusual behavior. On the other hand, due to complex dependencies between factors that can influence properties of goods, it is difficult to define in advance all possible problems (situations of interest) that can appear in the food logistics chain. For example, due to nonlinear relations between some environmental parameters (e.g. temperature and humidity) it is not possible to define complex crisp rules for expressing problem situations that will be observed in real-time. It means that the monitoring system should support data-analytics approach that will (dynamically) learn patterns from the past data in order to define situations to be monitored/detected in real-time. These situations usually correspond to unusual/anomalous behavior, which is main indicator that something unplanned (wrong) has happened.
4.2 High-level view on Complex Events Processing

The analysis from previous subsections requires so called closed-loop analytics (http://arstechnica.com/informationtechnology/2015/03/24/machine-consciousness-big-data-analytics-and-the-internet-of-things/) in transportation, an innovative form of mobile-client – server communication where local processing on a mobile device (truck) is driven by the complex processing performed on the server that includes (predictive) analytics based on the sensor data.

Following Figure 15 illustrate the complexity of the processing (real-time) data in the transportation domain.

![Figure 15: Complexity of the processing data in the transportation domain and the role of CEP](image)

Therefore, CEP enables remote connection of managers and drivers/vehicles in a proactive and personalized way, trying (a) to avoid some problem situations that the goods/vehicle/driver can be involved in and (b) to support the realization of some dynamic requests posted by customers/system. The situations of interest can be arbitrary complex, since we use complex event processing as the underlying technology for the real-time situational description. A source of information is the data sensed from the container (like information about location, temperature and humidity). For example, a supplier of a product could attach a tag to a pallet of fruit. The tags will repeatedly transmit temperature readings to a gateway unit installed on the refrigerated trailer (or railcar or container). In general case, the gateway for the communication
is the driver’s mobile phone. The gateway would then forward the temperature data to a server at the shipper headquarters or at the logistics provider, along with GPS location data.

The information to be transmitted can include personal driver data (for the case of security / using smartphone sensors or wearables) and localized environmental data (e.g. using open weather data). The system will be designed so that it can be easily extended with any other types of data sources.

This approach could also be used as a predictive maintenance tool to warn of equipment failure, to improve energy management, or to provide automatic record-keeping for regulatory compliance.

The usage of the system requires a web portal, where participating managers can define the requests (patterns: situations of interest) that will be deployed on a mobile phone of a selected driver/vehicle. In the case that situation of interest happen, the mobile app will send the alarm to the server, which will forward it with additional contextualization to the manager. In that way the vehicle and goods will be continuously monitored to enable a prompt reaction in the case of any issues, based on above mentioned data. Beside the real-time alarming support, the system will provide different types of statistics about vehicle/driver behaviour and any unusualness detected in the selected period (e.g. last week) will be delivered. This can help in better understanding the problems (if any) in the transportation process. Finally, the system will enable the collection and interchange between the patterns and data from different vehicles, leading to building stable communities which will boost the usage and the success of the system/approach.

The users will be able to define own situations of interest by using an editor, or by reusing the patterns from others (if allowed). Moreover, the system will provide a recommendation service that will recommend to managers some existing patterns based on their preferences or based on the behavior of the drivers or other users (additional service). Additionally and very important for the user acceptance: the most sophisticated security approaches will be introduced in order to ensure the prevention of data manipulation. The security and privacy will be based on our work in the eHealth domain, which is very demanding in this context.

Following Figure 16 illustrates the sequence of activities during the operation of the proposed system.
1. A manager is defining some patterns related to the situations she/he would be informed about (MonitoringRequest1)
2. The vehicle is getting this request as an automatically deployed pattern on app (MonitoringTask1)
3. The system is collecting all possible information from additional sources like social media, open data to update the detection of situations
4. In the same time the system will notify the driver if there is an issue with her/his vehicle
5. The app is collecting mobility data and processing locally in order to detect situations of interest in advance
6. This alarm is extended with contextual information and sent to the manager
7. The manager gets also some statistic information about the behaviour of the vehicle in order to get a better overview
8. The system can collect information from vehicle about any issue that happened
9. The system can also collect information from managers about their experiences
10. The system can suggest some new services based on the analytics of data provided so far
4.3 Benefits

In general, we argue that this approach will lead to an increased real-time situational awareness and faster, more accurate, real time decision making based on real-time data, that can make the re-routing process more efficient. In general case the increased situational awareness in the transportation domain will 1) enable avoiding problems before they happen and 2) discover opportunities proactively (therefore increase efficiency and reduce costs).

Examples of KPIs that can be optimized: shorter reaction time in the re-routing process, including reacting ahead of time, better understanding/discovery of the problems in the traffic, better precision in issuing alarms (less false positive alarms).

In following we elaborate on specific topics the customers can benefit on:

1. Enabling continuous data visibility/monitoring
   Our approach provides full visibility inside the transportation process, helping to ensure that users/travellers reach their destination under anticipated conditions. It is an emerging transportation model in which multiple sensors transmit data to multiple partners, enabling them to collaborate and respond to unexpected situations. The sensor devices detect current vehicle location and environmental variables such as temperature, light exposure, relative humidity, and barometric pressure, then wirelessly communicate these variables to the managers/customers.
   Apart from making complex tasks easier to plan and manage, transparency along the entire transportation process chain is ensured by using these solutions.

2. Keeping stakeholders informed
   Using geo-localization technology, our approach allows customers/managers to create customized, location-based alerts called geofences to inform them if the transportation process is going properly. This will enable a more efficient tracking and corresponding decision making.
   In general context, a range of industries can benefit from this type of monitoring. For example, healthcare and pharmaceutical companies shipping items that are either very difficult or impossible to replace—such as tissue specimens for clinical trials and/or diagnostic testing—need the environmental variables of these packages carefully maintained while in transit. By monitoring factors such as light exposure, temperature, and humidity, this solution could save the specimens from damage.
   Our approach gives users the control and insight needed to make better business decisions and stay ahead of the competition in a fast-paced and ever-changing marketplace. This, in turn, can lead to a superior customer experience and higher profit margin.

3. Decreasing reaction time – enabling proactivity
By receiving continuous data in the transportation process, users can take advantage of intervention services that will happen ahead of time (proactively), incl. finding alternate routes or inspecting and repackaging damaged goods. Instead of reacting to supply chain faults, our approach places control in the hands of the management team to proactively avoid these issues.

Even with the best of efforts, it’s impossible to eliminate all of the variables in the supply chain. Delays occur, and equipment breaks. But, by proactively monitoring the pallet-level temperature of fresh, frozen, and packaged foods, growers, producers, shippers, and retailers can not only more effectively manage the quality and safety of products as they move through the supply chain, but they can also document quality at delivery.

4. Supporting extended collaboration

Our approach supports very efficient, real-time collaboration between different stakeholders in the transportation process. Indeed, CEP enables: 1) better information gathering (more data sources can be used), 2) faster information integration (more efficient correlation can be done) and 3) more relevant alarming (more precise patterns/alarm situation can be defined)
5 Real time analytics

5.1 QMiner

Real time analytics demands fast and scalable high throughput analytics methods, that are able to cope with the data feeds in real-time. For this purpose, Optimum project will use extendable real time analytics platform - QMiner (JSI, n.d.) (http://qminer.ijs.si/). QMiner is an open-source data analytics platform for processing large-scale real-time streams containing structured or unstructured data. It is actively being used and developed by multiple research and commercial projects solving a numerous number of machine learning and stream analytics problems. These are ranging from anomaly detection, traffic prediction, recommendation systems and social sentiment analysis to the energy consumption prediction. It implements a comprehensive set of techniques for supervised, unsupervised and active learning on streams of data. It enables easy extraction of rich feature vectors from data streams using the data importing, normalization, re-sampling, merging and enrichment functionality. It also provides a range of online analysis methods, like exponential moving average, moving correlation, moving covariance, resampler, simple linear regression and others. QMiner platform presentation in context of Optimum is depicted in Figure 17.

The core of the QMiner platform and its components are written in C++ in order to achieve good performance. These native and fast methods are then exposed to the developers as JavaScript objects, which can be used in the Node.js scripting engine.

More detailed description of QMiner can be found in D1.4 (Conceptual Architecture).
Figure 17: QMiner.
5.1.1 Mobility patterns detection service

Mobility patterns detection and prediction is one of the analytics pipelines that can be built on top of QMiner. The pipeline consists of data cleaning, stay point detection algorithm, location clustering algorithm and the Markov Chain models for the prediction. The initial exploration in this direction started already at the end of Mobis EU project. There we identified the suitable algorithms that work online and are fast to develop. The very first version of the service was used to help with the personalization of mobility services. It collects knowledge about personal mobility patterns of users and enables us to provide users with more accurate notifications and recommendations without the need of specifying and saving all their commuting and travel paths. The very first prototype is described in D4.4 Personalized Mobility Patterns Learning (MobiS, 2015). However, it is also worth mentioning that a lot of further work was already done in scope of Optimum, including persistent storage of the history, evaluation methods, group analytics, integration with Optimum data-sources and libraries for mobile applications. This development has a twofold goal. On one side, it will provide real-time personal patterns analytics support to Optimum, while on the other side, it will allow us to evaluate the current algorithms and improve them, or switch with better versions when applicable. One such improvement (see Table 7) was already done during the scope of the Optimum project.

In order to acquire personalized mobility patterns, we have to obtain the physical locations of a person’s places and typical routes that matter to his/her daily life and routines. This is handled mainly by the GPS tracking services. These raw coordinates go through a few-steps process, from staypoint and path detection, to model building. We use term staypoint to refer to a location, where user stayed for a significant amount of time – e.g. home or work. The last step and the main goal is a prediction functionality, but several other features turned out to be interesting, e.g. collective analysis and derived recommendation system.
Core service

In order to be able to detect and predict user’s mobility patterns, we have to go through many quite independent processing steps (Figure 18). Each of these steps represents one layer of the mobility patterns detection and thus can be independently used as a service or an application.

1. Data collection (locations, places).
Several data sources can be used. We developed mobile app libraries for iOS and Android that send GPS data to server (described later), which we also plan to use in Optimum app.

We also collect data from Adria Mobil motorhomes which are equipped with devices that send GPS signal. Basically the algorithms are already robust enough, so any kind GPS data works.

On the server side, there is a simple API (RESTful service), where the client pushes the locations via the POST HTTP request. The server then sends the received locations through the steps 2 to 4.

2. Data cleaning and fusion.
This step involves removal of the obvious corrupted data, and time fusion of the sources.

3. Staypoint detection.
This is the first processing step, which processes the raw GPS coordinates and detects when user stays at some location for some non-trivial time.

Figure 18: Steps towards Personalized Mobility Patterns
The first algorithm that the data goes through is the Staypoint Detection, where the *staypoint* is defined as the limited area at which user stayed over a certain time interval. The algorithm that we currently use is an improved version of the Staypoint Detection (SPD) algorithm (Li, et al., 2008). The Staypoint detection service that is currently implemented works in real time, which means that immediately when it receives the location, it updates its models and replies with the detection results.

Because all of the detected activities are indexed and stored into a database, it is possible for the users to browse their travel and Staypoint history.

**Route detection.** In parallel with the Staypoint detection, we detect the routes between the Staypoints.

**Road-section mapping.** The routes detected in the previous step are usually taken on some road. This step takes the raw coordinates of the routes and maps them to the road network.

4. *Frequent Locations detection.*

As a result of the step 3, the Staypoints are being clustered and counted, to be able to separate more important and frequent locations from the less frequent ones.

Once the Staypoints and paths are stored into an index, the processing goes to the algorithm which is able to cluster the detected Staypoints. These are snapped together into clusters which correspond to frequent locations that can be analysed with machine learning and statistical algorithms. The implemented algorithm is online, which means that it processes new data when it comes in. Clusters use moving average for their aggregated data, meaning that centroid location can slightly change when new Staypoints are coming in. When two clusters are close enough to each other, the algorithm will merge them into one. However, issue arises when two locations are close by (one shop just next door from another etc.) and the algorithm does not differentiate and treats them as one location. It is planned as part of future work, to enable splitting such location into two locations.

5. *Model building and prediction.*

This step uses the data provided by the previous steps, to calculate the probabilities and distributions of the locations and visits. Using QMiner, the process then proceeds to modelling and prediction. In this step, we wanted to predict to which location the user will commute from current location, taking into account how much time she/he had spent at that location. From many possible approaches we chose to implement the first order discrete Markov Chain for initial version for the locations, and average stay-times before the jumps to each location, to assess the time of leaving the current location. Based on that model and the current user’s location and status, we are able to predict user’s next step.
## 5.1.2 Current state

### WEB GUI

Each user can access visualization of his/her data in a WEB GUI (real-time example: [http://traffic.ijs.si/NextPin/?user=TT1410](http://traffic.ijs.si/NextPin/?user=TT1410)). A screenshot of WEB GUI can be also seen in Figure 19. On right side, there is a list of all user’s activities detected as either staypoints (pin icon) or paths (man icon). When selecting a staypoint from that list, all other staypoints that correspond to the same frequent location are highlighted in red. On left side, another bar opens with information about this location – location ID, number of visits, visit times, suggestion what this location might be and histograms of visit frequency.

![WEB GUI Screenshot](image)

**Figure 19: Detected frequent location.**

### Location detection

Frequent locations are defined by coordinates which then need to be mapped to real locations like work, home, supermarket etc. Currently we use Foursquare to obtain venue suggestions around location. Then we run these venues through an algorithm which takes into account distance from location, type of activity, time of day etc., to select the most appropriate suggestion. In the example in figure 19 algorithm (correctly) suggested JSI building.
**Mobile (collection) apps**

We developed mobile apps for Android and iOS that send geolocation data to our server. They implement OS supported alarms, which wake up the phone every 30 seconds. After the phone is woken up, it checks whether there is recent good quality location data available. The collected location is then sent to the server, if the phone is online. If not, it is stored on the phone, and then sent in a bunch, when the phone comes online. The task proved to be a lot more difficult as expected since the latest OS enable less control to developers when working with alarms than older versions.

**Algorithms**

When considering the proper combination of methods and algorithms for our goal, we tested many approaches. It is important for our purpose that algorithm takes into account spatiotemporal information and handles well the noise. In Table 6 is presented comparison of several clustering algorithms that we considered. At the end we picked the solution that gives the best results given how demanding is regarding the CPU and memory, and regarding the development time.

**Table 6: Algorithms**

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Worst case running time</th>
<th>Spatial-temporal</th>
<th>Noise sensitivity</th>
<th>Returning staypoints/paths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Density Based Spatial Clustering of Application with Noise (Ester, Kriegel, Sander, &amp; Xu, 1996)</td>
<td>O(n^2)</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Spatial Temporal Density Based Spatial Clustering of Application with Noise (Birant &amp; Alp, 2007)</td>
<td>O(n^2)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Clustering-Based Stops and Moves of Trajectory (Alvares, et al., 2007)</td>
<td>O(n^2)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Stay Point Detection (Li, et al., 2008)</td>
<td>O(n^2)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ordering Points to Identify the Clustering Structure (Ankerst, Breunig, Kriegel, &amp; Sander, 1999)</td>
<td>O(n log n)</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
</tbody>
</table>

For implementation and evaluation, we considered two staypoint detection algorithms, both using time, longitude and latitude as input parameters:
1. **Traditional SPD**

Traditional staypoint detection works in a very basic way and is based on time and distance thresholds, and is described in (Li, et al., 2008). If distance is under a certain value and time is above a certain value, for example, distance equals 50 meters and time 30 minutes, we treat these points as staypoints, if they exceed these values, we treat them as a path.

2. **SPD with Two Passes**

This is an improved version of Traditional SPD algorithm and is the algorithm that we currently use. The first step is the same as for Traditional SPD, then additional step is made using results from the first step. This means that if needed, the stay points and path are merged in one single stay point if they exceed the thresholds. Stay points and paths are merged according to the following criteria:

- If two or three stay points are very close to each other, they will be merged into one stay point.
- If the algorithm detects a stay point at the first step, then a path and then a stay point, we check if the trajectory lasted less than a certain time and if the distance of the trajectory was shorter than a certain distance. If the calculated value is under a certain threshold, it will merge the stay points into one.

The result type is the same for traditional SPD and SPD with Two Passes.

If a certain activity is identified as a stay point, then the result will be a point, if it is identified as path, the result is a group of points which form a path.

**Evaluation**

We emphasize the evaluation of algorithms for the detection of stay points as it is an important process which tells us how well we calculated a certain activity. Our approach is different to other developers. Our criteria for determining the accuracy of a stay point considers:

- Start time
- End time
- The difference between the actual and calculated value.
- Type of activity - stay point or path

In addition to standard evaluation measures (Mean Absolute Error (MAE), Median Absolute Error (MdAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), Coefficient of determination ($R^2$), Explained Variance Score (EV)) we also used a non-standard measure, Normalized Cumulative Gain. It works according to the following procedure: if the SPD algorithm makes no mistakes, the one of
result (stay point or path) will collect 10 points. Firstly, it checks if the algorithm has accurately
determined the type, that is, whether it is a path/trajectory or a stay point. Incorrect type
detection is penalized severely. If the stay point algorithm’s result is same as the true value, it
will continue checking the difference in distance between the real and calculated location. If
differences in distance between true and calculated location exist, it will appropriately penalize
the result. It will then check the difference between real time, start time and end time.
Differences are appropriately penalized.

We labelled the data, calculated activities using Traditional SPD and SPD with Two Passes and
finally, ran a script with metrics to evaluate stay point detection algorithms.

Table 7: Algorithm result overview

<table>
<thead>
<tr>
<th></th>
<th>Traditional SPD</th>
<th>SPD with Two Passes</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG</td>
<td>360</td>
<td>1031</td>
</tr>
<tr>
<td>NCG</td>
<td>0.252</td>
<td>0.722</td>
</tr>
<tr>
<td>MAE</td>
<td>0.73</td>
<td>0.28</td>
</tr>
<tr>
<td>MdAE</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>MSE</td>
<td>0.7</td>
<td>0.24</td>
</tr>
<tr>
<td>MAPE</td>
<td>70.82</td>
<td>33.87</td>
</tr>
<tr>
<td>R2</td>
<td>-2.19</td>
<td>0.02</td>
</tr>
<tr>
<td>EV</td>
<td>-1.98</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Results for success at calculating stay points are given in Table 7. As we can see, SPD with Two
Passes is significantly better than traditional SPD. The best result for the algorithm SPD with Two
Passes is in bold in the table, and all measures assessed using the algorithm SPD with Two
Passes are better. The higher the score, the better is the model for measures NCG, EV in R2. The
lower the score, the better is the model that for measures MAE, MdAE, MSE, MAPE.

Normalized Cumulative Gain performs very well and carries the informative value, while we had
difficulties interpreting results for measures such as MAE and MSE. Normalized Cumulative Gain
could collect 1440 points for our data, Traditional SPD collected 360 and SPD with Two Passes
collected 1031. SPD with Two Passes assessed by NCG is 0.772, meaning that SPD with Two
Passes accurately calculated stay points or paths 77.2% of the time.
5.1.3 Next steps

Even though the algorithm works well for now, there is still a lot of space for improvement. To increase performance, we plan to change implementation of our algorithm from JavaScript to C++.

There are several things still to improve in staypoint detection, which is the first and most crucial step in the pipeline of patterns detection. One of them is to take into account accuracy of GPS location which would enable us to detect staypoints more accurately. Also related to accuracy is an issue, when algorithm merges two staypoints into a single one. We would like to enable users to split incorrectly merged location and then automatically re-cluster new locations. Now, several users may have each their own frequent location which actually corresponds to the same location. Natural next step would be to introduce global frequent locations which would be connected to user’s individual locations.

To improve predictions, we will upgrade first order- to second order Markov Chain and start using time as one of the features. Additionally, besides the staypoints, we plan to start treating paths as road sections, since now they are just sets of coordinates and we’re not extracting any information from them. This will allow us to do the predictions and behaviour analysis while on route as well.
6 References


20. Eleni I. Vlahogianni, Matthew G. Karlaftis, John C. Golias, Short-term traffic forecasting: Where we are and where we’re going, Transportation Research Part C 43 (2014) 3–19


32. JSI, n.d. QMiner. [Online]


39. Li, Q. et al., 2008. Mining user similarity based on location history. Reaserch Microsoft, p. 10.


61. Technol. 5, 1, Article 8.


69. Vij, Akshay; Carrel, André;


