TITLE OF THE RESEARCH THESIS PROPOSAL: Soft Computing and Big Data Mining techniques for reliable travel time prediction in urban environments

DIRS DOCTORATE PROGRAMME : Engineering for the Information Society and Sustainable Development

HOST RESEARCH UNIT (website link): Mobility Unit (http://research.mobility.deustotech.eu/)

BRIEF DESCRIPTION OF THE INTEREST OF THE RESEARCH THESIS PROPOSAL:

New patterns and models of business, commerce or leisure have **increased the demand for reliable transport systems**. This fact is **claimed by different reports and projects from national and international organizations**, such as the OECD (OECD 2010), the Transportation Research Board of United States (Transportation Research Board 2015), the Institute of Transport Studies of United Kingdom (Institute of Transport Studies 2013), or the European Road Transport Research Advisory Council (ERTRAC 2013). The offshore outsourcing of production, the adoption of just-in-time distribution systems, the tight scheduling of personal and freight activities, or the broadening of international trade are just some of the reasons behind this increasing importance of transport reliability.

According to the OECD report "Improving reliability on Surface Transport Networks" (OECD 2010), **one of the main policy options to improve transport reliability is the development of new Intelligent Transport Systems**, and more concretely, of Advanced Traffic Information Systems. These information systems should **provide the expected travel times and their possible variability to allow the users planning ahead**. They can contribute to reduce the stress caused by unreliability, mitigate the problems associated with delays, improve the service quality perceived by the user, or support traffic managers in the decision making process.

Travel Time Prediction Systems (TTPS) have arisen as a powerful tool for this purpose. Among the approaches proposed in literature, we can find methods based on Traffic Flow Theory (Celikoglu 2013; Li et al. 2013), on Data Mining (Simroth & Zahle 2011; Yildirimoglu & Geroliminis 2013), or on a hybridization of both (Li & Chen 2014; Elhenawy et al. 2014). Although travel time prediction systems has been widely studied by the scientific community (van Hinsbergen et al. 2007; Mori et al. 2014), **there are still some gaps to bridge in order to develop more accurate and reliable systems**. Some the most important are the next (Mori et al. 2014):

a. **Fusion of different data sources with traffic and context information:** The fusion of information from different type of sensors (loop detectors, interval detectors, etc.) or from different data sources (traffic operators, drivers, etc.) would lead to richer prediction models. Apart from traffic conditions, there are other factors that can influence travel time, as meteorology, special events, weekday, time of the day, etc. However, the use of context information has been hardly explored despite the strong impact of these factors on the traffic state. However, the fusion of information in TTPSs is still in its childhood as referred in (Mori et al. 2014).

b. **Inclusion of urban planning, social, demographic and economic factors in travel time prediction models.** It has been widely study that factors as land use, population density, urban form, accessibility or connectivity of the area, have an important influence in commuting behavior, and therefore travel time patterns (Antipova et al. 2011; Vojnovic et al. 2013). For example, industrialized urban areas are more congested on working days while leisure areas on holidays, areas that concentrate students have different travel patterns than those that concentrate

workers or families, etc. This type of information has been virtually ignored in the travel prediction models proposed in literature (Mori et al. 2014).

The emergence of **Intelligent Transport Systems (Perallos et al. 2015), Open Data and Volunteered Geographic Information has promoted the availability of a vast amount of data related to traffic, context, demography, economy, network infrastructure, etc. in urban areas**. Furthermore, many of these data are periodically updated, even in real time. The huge volume and variability of data available represents a valuable asset to develop better TTPS. However, it also **poses a real challenge for most of the existing techniques. This problem is commonly named with the term of Big Data (Lynch 2008)**, which refers to the difficulties and disadvantages of processing and analyzing huge amounts of data. Standard technologies and models do not provide users timely, cost-effective and quality predictions. Recent advances on Cloud Computing technologies allow us to adapt such techniques to be successfully applied over this changing and expanding mass data. Thus, **a new class of scalable TTPS that embraces the huge storage and processing capacities available nowadays in cloud platforms is required.**

MAIN RESEARCH AIMS:

The research proposed here is found on the interface between Intelligent Transportation Systems, Data Science and Economic Geography. **The goal of the thesis consists of developing more accurate and reliable travel time prediction systems using a multidisciplinariety approach that address the two issues mentioned above by means of Data Mining** (Han et al. 2011; Triguero et al. 2012), **Soft Computing** (Verdegay et al. 2008; Masegosa et al. 2014) **, Big Data technologies** (Lynch 2008; Triguero et al. 2015) and **Economic Geography methodologies** (Clark et al. 2003).

Data Mining techniques can be seen as tools for the discovery and analysis of knowledge from observed data. The main advantage of these methods versus other approaches for travel time prediction is the lower requirement of traffic theory knowledge. Soft Computing comprises a complete set of techniques, such as fuzzy systems, neural networks or metaheuristics, which have demonstrated its usefulness when applied in data mining problems (Zhang et al. 2014; Lopez-Garcia et al. 2015; Triguero et al. 2011). Especially when these techniques are hybridized and when the information processed presents uncertainty, imprecision or vagueness. Nevertheless, the application of standard Data Mining and Soft Computing techniques in large-scale data sets is not straightforward. The knowledge extraction process from big data has become a very difficult task for most of the classical and advanced Data Mining and Soft Computing tools. For this reason, the adaptation of data mining techniques to the emerging big data technologies is a must in order to overcome these limitations and make them scalable (Triguero et al. 2015; López et al. 2014). On the other hand, economic geography methodologies aim at studying the influence of the environment in human economic activities (Clark et al. 2003). Its application to urban environments will allow us to know how different aspects of the location and distribution of land use, economic activities, etc. intervene in the travel behavior of commuters. The inclusion of this information in travel time prediction systems could suppose a breakthrough in this field, since they have not been considered before.

Having said that, the **particular objectives of this thesis** will be the following:

- **A theoretical and empirical study of state-of-the-art techniques for travel time**

prediction in order to categorize existing trends and discover strengths and weaknesses of each family of methods.

- **Identification of urban planning, social, demographic and economic factors that influence commuting behavior,** and therefore, that have a direct impact on travel times.
- **The collection and analysis of different data sources with traffic, contextual, land use, social, demographic and economic information** (loop sensors, interval sensors, GPS, incidents, accidents, infrastructure, meteorology, population density, occupation, etc.).
- **Development of more accurate TTPS by means of Soft Computing-based Data Mining techniques and data fusion**, as Evolutionary Machine Learning (Fernández et al. 2015; Zhang et al. 2014; Lopez-Garcia et al. 2015; Triguero et al. 2011). These methods will fuse information from the different data sources collected.
- **Parallelization of the developed TTPS to enhance their scalability**. Concretely, we will rely on the success of the MapReduce paradigm (Dean & Ghemawat 2008), and beyond it with Apache Spark (Zaharia et al. 2012).

This thesis will be framed within the H2020 project:

TIMON: Enhanced real time services for optimized multimodal mobility relying on cooperative networks and open data. (http://www.timon-project.eu/)

This project is a **consortium of 11 organizations, formed by Universities, RTOs, SME and big industry from 8 different European countries** (Spain, Italy, Belgium, the United Kingdom, Germany, Hungary, the Netherlands, and Slovenia). Its objective consists of addressing problems related to congestion, traffic safety and environmental challenges by creating a cooperative ecosystem where people, vehicles, infrastructure and businesses are connected. **This project will provide the data and infrastructure required by the MSC Researcher to successfully reach all the objectives pursued in this thesis.**

WORK PLAN

This research proposal is formed by five different work packages corresponding to each of the objectives. These work packages and tasks comprise the 36 months work plan for the MSC researcher.

- **WP1: Theoretical and empirical study of state-of-the-art techniques for travel time prediction (M1-M6)**

    - **T1.1 Study and categorization of the state-of-the-art in travel time prediction (M1-M3):** This task aims at reviewing the specialized literature, identifying the most relevant works, different categories of methods, trends, as well as strengths and weaknesses of state-of-the-art methods.

    - **T1.2 Study and categorization of data sources and data sets used in literature (M1-M3):** In parallel with T1.1, the MSC researcher will identify the most common traffic and contextual data sources and data sets used in literature, as well as others publicly available. The outcome of this task will be a set of benchmarks for travel time prediction.

- **T1.3 Implementation and Experimentation (M3-M6):** The objective of this task consist of implementing an open-source module containing most of the current approaches and all the necessary tools to reproduce the experiments carried out in the research papers. These experiments will be performed over the benchmark created in task T1.2. This task will allow us to characterize, from an empirical point of view, the performance of each method according to its family.

- **WP2: Analysis and identification of urban planning, social, demographic and economic factors that influence commuting behavior in urban environments (M4-M10).**

    - **T2.1 Review and categorization of studies related to the influence of these factors in commuting or travelling behavior (M4-M8).** With the collaboration of Dra. Montserrat Pallares-Barbera, the student will review and categorize the studies in economic geography that relates urban planning, social, demographic and economic factors to commuting behavior in urban environments

    - **T2.2 Identification of most relevant factors influencing travelling behavior (M7-10).** After the review of the former studies, the student will identify the most relevant factors that will be used in the next stages of the project in order to enrich the variability and significance of the information available to develop travel time prediction systems.

- **WP3: Collection and analysis of different data sources with traffic and contextual information (M10-M17)**

    - **T3.1 Identification and classification of new relevant data sources for traffic and contextual data (M10-M11):** The student will identify which data sources, apart from the commonly used in literature, could provide useful information for travel time prediction, focusing specially on contextual data. Most of this data will be provided by the H2020 project TIMON.

    - **T3.2 Preparation of datasets (M11-M12):** Once the relevant data sources have been identified and categorized, the next step will be to create datasets that integrates all this information in a proper format. The datasets developed here will be the benchmark that will be used to test the proposals of this thesis.

    - **T3.3 Data preprocessing (M12-M17):** This task aims at preprocessing data by applying instance reduction and feature selection techniques.

- **WP4: Development of more accurate TTPSs by means of fusion of different data sources (M18-M26)**

    - **T4.1 New Soft Computing-based Data Mining techniques for travel time prediction (M18-M22):** This task will focus on the design of new Data Mining algorithms based on Soft Computing techniques that allow us to fuse information from the different data sources, identified in the previous work

package, to improve the accuracy of state-of-the-art methods.

- **T4.2 Implementation, experimentation and validation (M22-M26):** The objective of this task is to implement the techniques developed in task T3.1, perform a wide battery of experiments according to a methodology previously established, and validate the performance of these techniques comparing them versus state-of-the-art algorithms.

- **WP5: Parallelization of TTPS (M26-M33):**

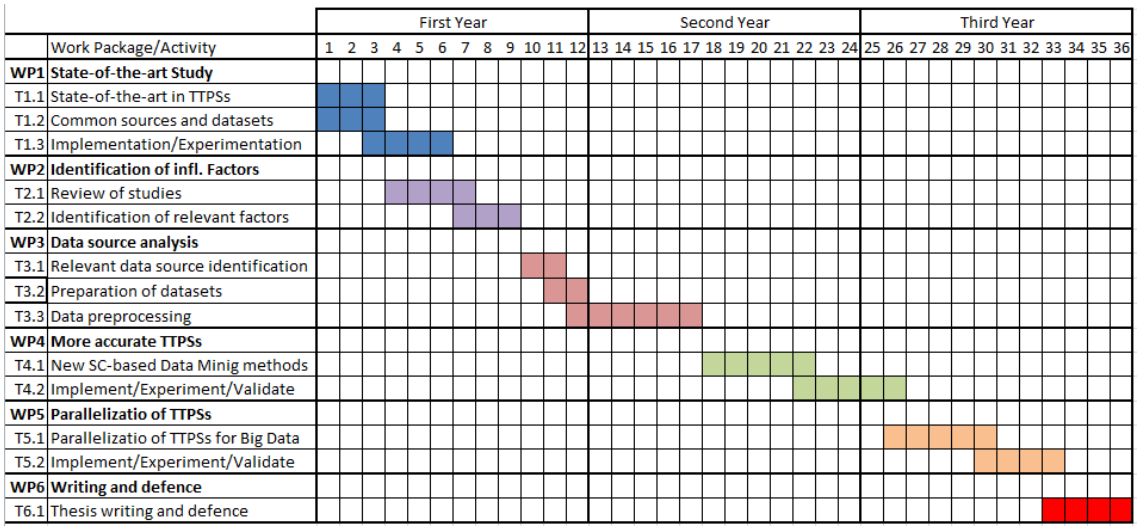
  - **T5.1 Parallelization of TTPS to address Big Data challenges (M26-M30).** This task aims at designing parallel approaches to speed up the travel time prediction process without degrading its accuracy or reliability. To this end, we will make use of Big Data technologies as Hadoop or Apache Spark.

  - **T5.2 Implementation, experimentation and validation (M31-M33):** This task is analogous to T4.2

- **WP6: Thesis writing and defense (M32-M36)**

  - **T6.1 Thesis manuscript writing and public defense (M32-M36)**

GANTT DIAGRAM



BIBLIOGRAPHY:

Antipova, A., Wang, F. & Wilmot, C., 2011. Urban land uses, socio-demographic attributes and commuting: A multilevel modeling approach. *Applied Geography*, 31(3), pp.1010–1018.

Celikoglu, H.B., 2013. Flow-Based Freeway Travel-Time Estimation: A Comparative Evaluation Within Dynamic Path Loading. *IEEE Transactions on Intelligent Transportation Systems*, 14(2), pp.772–781.

Clark, G.L. et al., 2003. The Oxford handbook of economic geography, Oxford University Press.

Dean, J. & Ghemawat, S., 2008. MapReduce: simplified data processing on large clusters.

Communications of the ACM, 51(1), pp.107–113.

Elhenawy, M., Chen, H. & Rakha, H.A., 2014. Dynamic travel time prediction using data clustering and genetic programming. *Transportation Research Part C: Emerging Technologies*, 42, pp.82–98.

ERTRAC, 2013. *Multi-Annual Implementation Plan for Horizon 2020*, Available at: http://www.ertrac.org/uploads/documentsearch/id20/ertrac-map-h2020_67.pdf.

Fernández, A. et al., 2015. Revisiting Evolutionary Fuzzy Systems: Taxonomy, applications, new trends and challenges. *Knowledge-Based Systems*, 80, pp.109–121.

Han, J., Kamber, M. & Pei, J., 2011. *Data Mining: Concepts and Techniques*

van Hinsbergen, C.P., van Lint, J.W.C. & Sanders, F.M., 2007. Short term traffic prediction models. In *Proc. of the 14th World Congress on Intelligent Transport System (CD-ROM)*.

Institute of Transport Studies, 2013. Modelling and Valuing Reliability and Punctuality. Available at: http://www.its.leeds.ac.uk/research/themes/reliability/.

Li, C.-S. & Chen, M.-C., 2014. A data mining based approach for travel time prediction in freeway with non-recurrent congestion. *Neurocomputing*, 133, pp.74–83.

Li, L. et al., 2013. Freeway Travel-Time Estimation Based on Temporal–Spatial Queueing Model. *IEEE Transactions on Intelligent Transportation Systems*, 14(3), pp.1536–1541.

López, V. et al., 2014. Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. *Fuzzy Sets and Systems*, 258, pp.5–38.

Lopez-Garcia, P. et al., 2015. A Hybrid Method for Short-Term Traffic Congestion Forecasting Using Genetic Algorithms and Cross Entropy. *IEEE Transactions on Intelligent Transportation Systems*, PP(99), pp.1–13.

Lynch, C., 2008. Big data: How do your data grow? *Nature*, 455(7209), pp.28–9.

Masegosa, A.D. et al., 2014. *Exploring Innovative and Successful Applications of Soft Computing* A. D. Masegosa et al., eds., IGI Global.

Mori, U. et al., 2014. A review of travel time estimation and forecasting for Advanced Traveller Information Systems. *Transportmetrica A: Transport Science*, 11(2), pp.119–157.

OECD, 2010. *Improving reliability on Surface transport networks*, Available at: http://internationaltransportforum.org/Pub/pdf/10Reliability.pdf.

Perallos, A. et al., 2015. *Intelligent Transport Systems: Technologies and Applications*, Chichester, UK: John Wiley & Sons, Ltd.

Simroth, A. & Zahle, H., 2011. Travel Time Prediction Using Floating Car Data Applied to Logistics Planning. *IEEE Transactions on Intelligent Transportation Systems*, 12(1), pp.243–253.

Transportation Research Board, 2015. Second Strategic Highway Research Program (SHRP 2). Available at: http://www.trb.org/Publications/PubsSHRP2ResearchReportsReliability.aspx.

Triguero, I. et al., 2012. A Taxonomy and Experimental Study on Prototype Generation for Nearest Neighbor Classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*

(Applications and Reviews), 42(1), pp.86–100.

Triguero, I. et al., 2015. MRPR: A MapReduce solution for prototype reduction in big data classification. *Neurocomputing*, 150, pp.331–345.

Triguero, I., García, S. & Herrera, F., 2011. Differential evolution for optimizing the positioning of prototypes in nearest neighbor classification. *Pattern Recognition*, 44(4), pp.901–916.

Verdegay, J.L., Yager, R.R. & Bonissone, P.P., 2008. On heuristics as a fundamental constituent of soft computing. *Fuzzy Sets and Systems*, 159(7), pp.846–855.

Villacorta, P.J. et al., 2014. A new fuzzy linguistic approach to qualitative Cross Impact Analysis. *Applied Soft Computing*, 24, pp.19–30.

Vojnovic, I. et al., 2013. The Burdens of Place: A Socio-economic and Ethnic/Racial Exploration into Urban Form, Accessibility and Travel Behaviour in the Lansing Capital Region, Michigan. *Journal of Urban Design*, 18(1), pp.1–35.

Yildirimoglu, M. & Geroliminis, N., 2013. Experienced travel time prediction for congested freeways. *Transportation Research Part B: Methodological*, 53, pp.45–63.

Zhang, X. et al., 2014. Hierarchical fuzzy rule-based system optimized with genetic algorithms for short term traffic congestion prediction. *Transportation Research Part C: Emerging Technologies*, 43, pp.127–142.

Zhang, X. & Rice, J.A., 2003. Short-term travel time prediction. *Transportation Research Part C: Emerging Technologies*, 11(3-4), pp.187–210.

Zaharia, M. et al., 2012. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. 9th USENIX conference on Networked Systems Design and Implementation pp. 1-14.

THESIS DIRECTOR(S):

Antonio D. Masegosa (co-director), IKERBASQUE Research Fellow, University of Deusto, Bilbao, Spain

Isaac Triguero (co-director), Assistant Professor, University of Nottingham, Nottingham, United Kingdom

Montserrat Pallares-Balles (external collaborator), Associate Professor, Autonomous University of Barcelona, Barcelona, Spain.

EXCELLENCE

5 Main academic publications related to the topic (impact factor or other quality assessment criteria):

1. PJ Villacorta, **A. D. Masegosa**, D Castellanos, MT Lamata, A new fuzzy linguistic approach to qualitative Cross Impact Analysis. Applied Soft Computing 24, 19-30, 2014, doi: 10.1016/j.asoc. 2013.08.006
   (JCR IF (2014): 2.810, Ranking: 17/123 (Q1), Google Scholar Citations: 6)

In this paper we proposed an extension of the so called MICMAC method, a cross impact analysis procedure widely used in scenario planning. Despite the success of this method, it still showed many drawbacks to capture and model the uncertainty present in this type of studies. In this work, we successfully solve these issues by incorporating linguistic variables and fuzzy sets to the original procedure. The results showed the superiority of our proposal.

2. **A. D. Masegosa**, D. Pelta, I. G. del Amo, The role of cardinality and neighborhood sampling strategy in agent-based cooperative strategies for Dynamic Optimization Problems. Applied Soft Computing, 14, Part C, 577-593, 2014 doi:10.1016/j.asoc.2013.08.006
   (JCR IF(2014): 2.810, Ranking: 17/123 (Q1), Google Scholar Citations: 6)

   This paper supposed a step forward to show the competitiveness of trajectory-based methods in Dynamic Optimization Problems. This contrasted with the common believe in this field. We proved that by designing a good cooperation scheme among trajectory-based methods run in parallel, they can improve the performance of state-of-the-art population-based algorithms for these problems.

3. **A. D. Masegosa**, D. A. Pelta, J. L. Verdegay. A centralised cooperative strategy for continuous optimisation: The influence of cooperation in performance and behavior. Information Sciences, 219, 573-92, 2013 doi:10.1016/j.ins.2012.07.002
   (JCR IF (2013): 3.893, Ranking: 8/135 (Q1), Google Scholar Citations: 7)

   This publication appeared in one of the most prestigious journals in our field. It is ranked in the first decile of its category and its JCR Impact Factor is 3.893. In this work, we used fuzzy rules to design a cooperative scheme among trajectory-based metaheuristics for continuous optimization problems. Similarly to the previous work, we showed that cooperative trajectory-based methods can obtain similar or better results than state-of-the-art algorithms.

4. **I. Triguero**, J. Derrac, S. García, F. Herrera. A taxonomy and experimental study on Prototype Generation for Nearest Neighbor Classification. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews 42 (1) (2012) 86-100, doi: 10.1109/TSMCC. 2010.2103939.
   (JCR IF (2012): 2.548, Ranking: 17/115 (Q1), Google Scholar Citations: 140)

   This work reviewed the state-of-the-art methods in prototype generation at the time of writing. It has been widely accepted in the specialized literature since it provided open-source software, and guidelines for practitioners about future work and gaps presented in the literature.

5. **I. Triguero**, S. García and F.Herrera, Differential Evolution for Optimizing the Positioning of Prototypes in Nearest Neighbor Classification. Pattern Recognition 44 (4) (2011) 901-916, doi: 10.1016/j.patcog.2010.10.020
   (JCR IF (2011): 2.292, Ranking: 18/111 (Q1), Google Scholar Citations: 79)

   We applied evolutionary algorithms to perform prototype generation in conjunction with prototype selection. The proposed hybrid technique showed to perform better than isolated methods. This algorithm has become the state-of-the-art in prototype reduction so far, being used in multiple applications by other authors.

6. **I. Triguero**, D. Peralta, J. Bacardit, S. García, F. Herrera. MRPR: A MapReduce Solution for Prototype Reduction in Big Data Classification. Neurocomputing 150 (2015), 331-345.
(JCR IF (2014): 2.083, Ranking: 36/123 (Q2), Google Scholar Citations: 57)

This is the first paper that performs prototype reduction in the context of Big Data based on the MapReduce approach. Despite being published this year, it has already caught the attention of a wide number of researchers.

A **complete list of publications** for both researchers can be found in the links below:

Antonio D. Masegosa: https://scholar.google.es/citations?user=xNqilwoAAAAJ
Isaac Triguero: https://scholar.google.es/citations?user=KogUP4YAAAAJ

The complete list of publications of the external collaborator can be found here:

Monserrat Pallares-Barbera: http://scholar.harvard.edu/montserrat-pallares-barbera/publications

Main projects developed by the directors related to the topic (especially european projects). It is very valuable that the research thesis proposal is aligned with the European projects approved that the research host unit has to develop.

1. **TIMON: Enhanced real time services for optimized multimodal mobility relying on cooperative networks and open data.**

   Funding Agency: Horizon 2020 - MG-3.5a-2014 - Cooperative ITS for safe, congestion-free and sustainable mobility [Grant Agreement – 636220]
   Founding: 5.605.213 €
   Period: June 2015- December 2018
   Principal Researcher: Asier Perallos
   Participant: Antonio D. Masegosa (Researcher)
   Web: http://www.timon-project.eu/

2. **ASCETAS: Applicability of Soft Computing in Technologically Advance Environments. Ref: TIN2011-27696-C02-01**

   Funding Agency: Spanish Ministry of Economy and Competitiveness, TIN2011-27696-C02-01
   Founding: 31.218 €
   Period: January 2012- December 2014
   Principal Researcher: José L. Verdegay
   Participant: Antonio D. Masegosa (Researcher)
   Web: http://modo.ugr.es/ASCETAS/

3. **Optimization Strategies in Intelligent Systems: Applications to Dynamic**

Environments

Funding Agency: Spanish Ministry of Science and Innovation, TIN2008-01948
Period: January 2009 - June 2012
Founding: 25.410 €
Principal Researcher: David Pelta
Participant: Antonio D. Masegosa (Researcher)
Web: http://www.dynamic-optimization.org/

4. **OPTIRAIL: Development of a Smart Framework based on Knowledge to Support Infrastructure Maintenance Decisions in Railway Networks**

   Funding Agency: Seventh European Framework (FP7) -SST.2012.5.2-2. - Next generation tools for optimised infrastructure asset management [Grant Agreement-314031]
   Funding: 3.916.343 €
   Period: January 2012 - September 2015
   Principal Researcher: J.M. Benítez
   Participant: Isaac Triguero (Researcher)
   Web: http://www.optirail.eu/

5. **APlaCA (Advanced Platform Cloud for Andalucía): Cloud Platform**

   Funding Agency: Research Contract between the company Indra Sistemas S.A. and the University of Granada
   Founding: 445.804 €
   Period: December 2011 - May 2015
   Principal Researcher: F.Herrera and J.M. Benítez
   Participant: Isaac Triguero (Researcher)

6. **La Palma Coop: Application of Computational Intelligence and Machine learning to the prediction of recollection performance and detection of plagues in greenhouses.**

   Funding Agency: Research Contract between the company La Palma Coop and the University of Granada
   Founding: 75.000 €
   Period: October 2013 - October 2014
   Principal Researcher: F.Herrera and J.M. Benítez
   Participant: Isaac Triguero (Researcher)

INTERDISCIPLINARIETY

Description of the different knowledge areas that the research proposal integrates :

The **research posed in this thesis is located on the intersection among three knowledge areas, Intelligent Transportation Systems, Data Science and Economic Geography**. Below, we briefly describe these areas:

- ***Intelligent Transport Systems:*** According to the EU Directive 2010/40/EU, Intelligent Transport Systems (ITSs) are advanced applications which without embodying intelligence as such aim to provide innovative services relating to different modes of transport and traffic management and enable various users to be better informed and make safer, more coordinated and 'smarter' use of transport networks. ITSs integrate telecommunications, electronics and information technologies with transport engineering.

- **Data Science:** This discipline can be seen as a group of principles to guide the extraction of information from data. Within the different knowledge areas and methodologies in Data Science, this project will be mainly focused on the next three:

    - *Data Mining:* This is the process of analyzing data from different perspectives and extracting relevant information, in an automatic way. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining uses advanced statistical tools to reveal trends, patterns, and relationships, which might otherwise have remained undetected. In contrast to an expert system (which draws inferences from the given data on the basis of a given set of rules) data mining attempts to discover hidden rules underlying the data.

    - *Soft Computing:* The term Soft Computing was firstly defined by Zadeh in 1994. Although there are different points of view, in general terms, Soft Computing can be understood as mix of different computational methods that present tolerance to uncertainty and imprecision, and which can be combined to obtain robust, manageable and low cost solutions. The main categories of computational methods that compose Soft Computing are Artificial Neural Networks, Probabilistic Reasoning, Fuzzy Systems and Metaheuristics.

    - *Big Data Learning:* It involves data whose volume, diversity and complexity requires new techniques, algorithms and analyses to extract valuable knowledge. The most well-known approach to handle these limitations is the MapReduce framework that offers a simple but robust environment to tackle the processing over a cluster of computing elements. Some of the most popular implementations of this framework are Apache Hadoop and the recent Apache Spark, which provides even more flexibility to tackle big data sets.

- ***Economic Geography:*** It aims at studying the influence of the environment in the economic activities of man and how they affect to their location, distribution and spatial organization. Economic geography it is a very broad discipline in which different methodologies, both economical and geographical, have been used to study various phenomena as location of industry, economies of agglomeration, transportation, real state, gentrification or the economics of urban form.

Other mechanisms different from co-direction to implement the interdisciplinariety:

The three main disciplines that combine this proposal are Data Science, Economic Geography and Intelligent Transportation Systems.

The background knowledge related to Data Science will be provided by the thesis directors, Antonio D. Masegosa and Isaac Triguero**. Dr. Masegosa[1], awarded as IKERBASQUE Research Fellow**, has a strong experience on Soft Computing, particularly on metaheuristics and fuzzy systems, fields in which he has been working since he started his PhD in 2006. During this period he has published three books, 15 JCR indexed papers, and more than 20 articles in both national and international conferences. **Dr. Triguero[2] is a promising young researcher that holds 22 JCR publications** related to Data Mining and Big Data, his main knowledge areas. Furthermore, he counts with more 20 contributions in national and international conferences. Apart from this, it is important to highlight that **both researchers have experience in PhD thesis mentoring**. Concretely, Dr. Masegosa has co-advised a PhD dissertation and is currently co-adving another one; and Dr. Triguero is the current co-advisor of a joint PhD thesis between the University of Nottingham (United Kingdom) and the University of Granada (Spain).

The required expertise for Economic Geography will be provided by Dra. Monserrat Pallares-Barbera[3], who will participate in this proposal as external collaborator. Dra. Pallares-Barbera is the **Head of the research group in Economic Geography of the Autonomous University of Barcelona (Spain)**. She has a wide trajectory in research, teaching and knowledge transfer. She has been **visiting scholar in various outstanding research centers** as a the Institute for Quantitative Social Science of the **Harvard University (USA)**, the department of geography of the **Boston University (USA)** and the Faculty of Geography of the **Clerk University (USA)**. She has published 13 JCR indexed papers, supervised 5 PhD theses and directed 9 research projects.

The required expertise about Intelligent Transport Systems will be provided by the Host Research Group, the Mobility Unit at DeustoTech. This group has excellent researchers in this field as Dr. Asier Perallos[4] and Dr. Enrique Onieva[5]. **Dr. Perallos was the former Principal Investigator of this unit and coordinator of the H2020 project TIMON**. He has lead more than a dozen of projects in this area and published nearly 20 JCR papers. In turn, **Dr. Onieva has also an extensive experience in this field where he accumulates more than 20 papers in JCR journals.**

The interdisciplinarity of this thesis will be also strengthened by the H2020 project TIMON. Apart from the knowledge area of this thesis proposal, this project involves fields as Data Harmonization, Open Data Management, Cooperative Positioning based on GNSS, Vehicular Hybrid Communications, Cloud Computing and Mobile Applications. Although these areas will not be covered in the thesis, the participation in this project will offer the MSC researcher a wide view of different fields and technologies related to Intelligent Transport Systems and how they can be used to implement safety and informational services for drivers, businesses and vulnerable road users.

INTERNATIONALIZATION

[1] https://scholar.google.es/citations?user=xNqilwoAAAAJ
[2] https://scholar.google.es/citations?user=KogUP4YAAAAJ&hl
[3] http://scholar.harvard.edu/montserrat-pallares-barbera/publications
[4] http://research.mobility.deustotech.eu/people/members/asier-perallos/
[5] https://scholar.google.es/citations?user=--6f3YEAAAAJ

Relationship with H2020 topics or ERC bottom-up initiatives:

The research proposed in this thesis is aligned with the **H2020 societal challenge on Smart, Green and Integrated Transport**, especially with two of its key objectives:

- ***Better mobility, less congestion, more safety and security.*** As it will be explained below, the research done in this thesis could contribute to reduce congestion and road traffic accidents.
- ***Global leadership for the European transport industry.*** The research proposed here would help to improve the design, planning and reliability of European distribution networks, and therefore to the global leadership of European transport industry.

In fact, this topic fits with many of the calls that will be open on the **H2020 Work Programme 2016-2017 for Smart, Green and Integrated Transport**. Some of this call are the following:

- MG-5.1-2016: Networked and efficient logistics clusters
- MG-5.2-2017: Innovative ICT solutions for future logistics operations
- MG-6.1-2016: Innovative concepts, systems and services towards 'mobility as a service'
- MG-8.5-2017: Shifting paradigms: Exploring the dynamics of individual preferences, behaviours and lifestyles influencing travel and mobility choices.

The topic of this thesis is also aligned with the **strategic research agenda of the European Road Transport Research Advisory Council (ERTRAC)**. It has included reliability as one of the three main societal needs of the European Transport Systems for 2030, together with decarbonization and safety[6]. Concretely, one of the specific targets established by this institution was to increase the reliability of transport schedules by a 50 percent by 2030.

International mention:
During the thesis, **we plan at least two short research stays**, from three to four months, **with the co-director Dr. Isaac Triguero, at the University of Nottingham, in United Kingdom.**

Partner university:
University of Nottingham (United Kingdom)

International co-direction or international "co-tutelle":
International co-direction between Dr. Antonio D. Masegosa, from the University of Deusto (Spain), and Dr. Isaac Triguero, from the University of Nottingham (United Kingdom).

Other mechanisms different from the previous to implement the interdisciplinariety:

Apart from international co-direction, **the participation in the TIMON project will foster the internationalization of the research done in this thesis**. The project consortium is formed by Universities, Research and Technology Organizations, Small and Medium Enterprises and big industry from eight different countries in Europe (Spain, Italy, Belgium, the United Kingdom, Germany, Hungary, the Netherlands, and Slovenia). In this way, the MSC researcher will have the opportunity for international networking with both academicians and professionals from different institutions across the mentioned countries.

[6] http://www.ertrac.org/uploads/documentsearch/id20/ertrac-map-h2020_67.pdf

Another planned mechanism to implement internationalization is the participation of the MSC Researcher in prestigious international conferences as IEEE Intelligent Transportation Systems Conference, the IEEE World Congress on Computational Intelligence, the Genetic and Evolutionary Computation Conference, etc.

INTERSECTORIALITY

Description of the social impact of the research:

As mentioned above, the main research aims of the thesis are to improve the accuracy of current TTPSs, by the fusion of different data sources with traffic and contextual data, on one hand, and to improve their reliability, by providing information about both the expected travel time and its variability. **These advances with respect to the state-of-the-art would have a high impact on road transportation systems and especially on those public and private services that depends on them** as logistics, supply chains, public transport, etc. Some of the main impacts are the following:

- ***Increase the Quality of Service (QoS) perceived by users of transportation systems.*** Many studies show reliability as a key factor in the quality of transportation system. Having available more accurate estimations of travel times and, even more important, information about how these times can oscillate, would increase users' satisfaction.
- ***Improve design, planning and reliability of distribution networks*****.** Travel times among hubs, customers, factories, providers, warehouses, etc. are pivotal in the design and planning of distribution networks. The research posed here will allow us to carry out these tasks using more precise and rich information about travel times, which it would result in a higher reliability.
- ***Facilitate the implementation of highly efficient production methodologies as Lean or Just-in-Time.*** Among other aspects, these methodologies rely on lowering warehousing costs by increasing delivering frequency and reducing the sizes of the transported slots. Decrease warehousing implies a higher confidence on deliveries being in time when they are required. Increasing the reliability of TTPSs would help to the implantation of these production methodologies.
- ***Reduce traffic congestion*****.** Routes passing through road stretches with a higher probability of congestion are usually subject to lower reliability in travel times. If users are better informed about these facts, they will tend to avoid these routes.
- ***Reduce traffic accidents.*** The stress motivated by a bad travel planning or unexpected road conditions (getting late to work, delays in deliveries, be at time in important appointments, etc.) is an important cause of road accidents. The systems developed in this thesis would help the users to do a better planning and to be aware of possible delays.

The high impact of this proposal is also supported by the interest of the company GEOX LTd (http://www.geox.hu/) **in the results of this thesis.** GEOX LTd. is a company whose main business areas are different services related to Geographical Information Systems. They have an extensive experience in research project at EU level, where they have participated in project as Sopcawind (http://www.sopcawind.eu/), Fusepool (http://www.fusepool.eu/) or

TIMON (http://timon-project.eu/).

Name of the social agent: GEOX Ltd. (http://www.geox.hu/)

GEOX Ltd agreed to host the MSC researcher for at least three weeks in their facilities in order to carry out different tasks in the frame of this project as studying the necessities that they have identified in terms of travel time prediction reliability.

Possibility of industrial doctorate title:
This thesis does not consider the possibility of an industrial doctorate title.

RELATED ON-GOING PROJECT

The PhD thesis proposal will be connected with the following research projects:
This thesis will be connected in first place with the H2020 project **TIMON: Enhanced real time services for optimized multimodal mobility relying on cooperative networks and open data. (http://www.timon-project.eu/)**

This project is a **consortium of 11 organizations, formed by Universities, RTOs, SME and big industry from 8 different European countries** (Spain, Italy, Belgium, the United Kingdom, Germany, Hungary, the Netherlands, and Slovenia). Its objective consists of addressing problems related to congestion, traffic safety and environmental challenges by creating a cooperative ecosystem where people, vehicles, infrastructure and businesses are connected. **This project will provide the data and infrastructure required by the MSC Researcher to successfully reach all the objectives pursued in this thesis.**

In second place, this PhD thesis will be aligned with another **H2020 proposal that is currently in preparation**. Its name is **LOGISTAR: Enhanced Data Management Techniques For Logistics Planning And Scheduling In Real Time.** The project is **a consortium of 12 organizations, formed by Universities, RTOs, SME and big industry from 9 different European countries** (Spain, Ireland, Servia, the United Kingdom, Germany, Italy, Belgium, Hungary and Austria). The main objective of the project is to allow effectively planning and optimizing transport operations in the logistic supply chain by taking advantage of horizontal and vertical collaboration as well as the data gathered from the interconnected environment through a real-time decision making tool developed with the purpose of delivering information and services to freight transport operators, their clients and other stakeholders such as warehouse or infrastructure managers

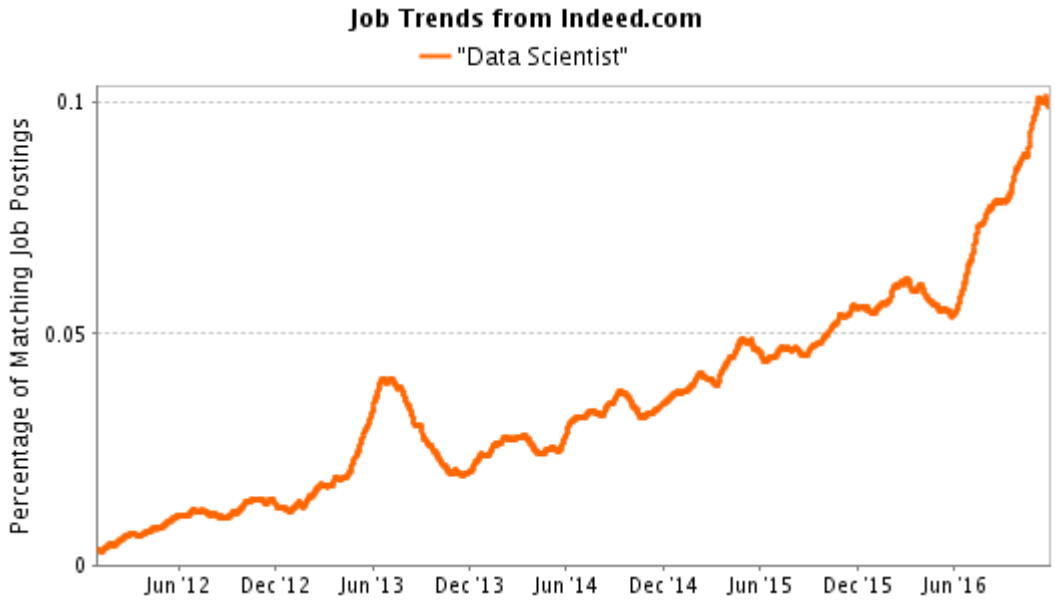
EMPLOYABILITY OR FUTURE CAREER POSIBILITIES RELATED TO THE RESEARCH POSITION:

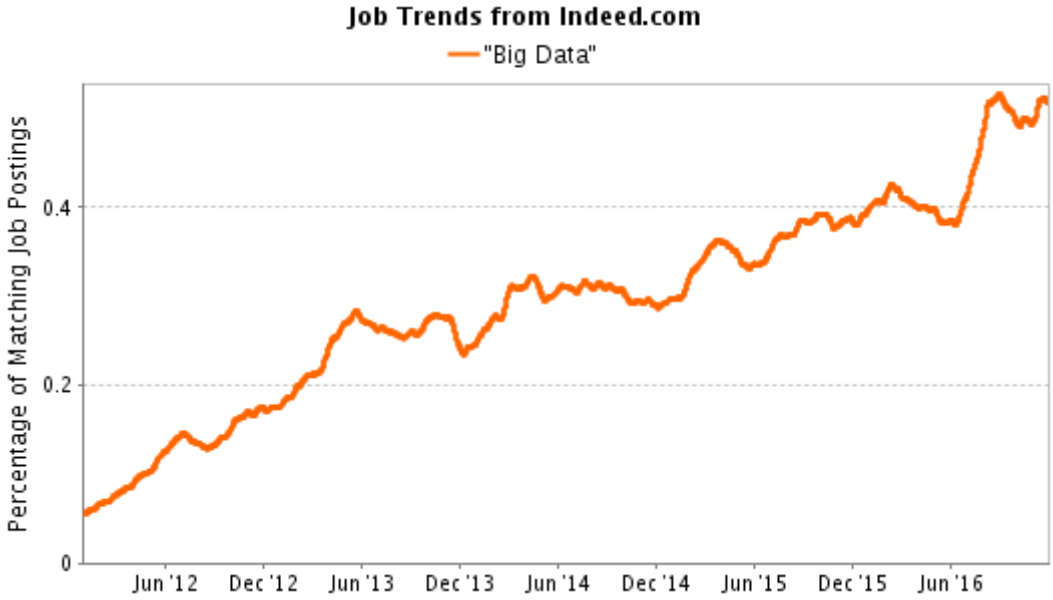
Along the thesis, **the PhD student will acquire important vertical and horizontal or transferable skills** that will improve her/his employability and hence, her/his future career possibilities.

Regarding vertical skills, the work addressed in this thesis will provide the MSC researcher the **necessary skills to perform advanced tasks related to Data Science.** Data science is **one of the most demanded skills worldwide**. The plot below shows the percentage of jobs postings that contain the word "Data Scientist" and "Big Data", and it has been extracted from Indeed.com[7]. The graphic clearly shows the increasing demand of this profile in the job market along last

[7] http://www.indeed.com/trendgraph/jobgraph.png?q=%22Data+Scientist%22

years. In fact, the **Harvard Business Review catalogued data scientist as "the sexiest job of the 21st century"**[8] in 2012.





The **main skills related to Data Science that the MSC researcher will develop** are:

- Data-driven problem solving
- Programming in languages or frameworks as R, Matlab, Java 8, etc.
- Big Data problem solving
- Big Data technologies as Hadoop, Spark, Flink, etc.
- Statistical analysis
- Data Mining and Machine learning methods for prediction
- Data visualization and communication.

The PhD student will also develop important transferable skills required by knowledge based

[8] https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

economy. Some of the most important are the following:

- **Communication skills:** the MSC Researcher will be encouraged to do periodical speeches to show their research advances and to participate in international conferences, forums, etc.

- **Teamwork skills:** the interdisciplinary and international character of the thesis proposed will require an important interaction and cooperation with different researchers.

- **Entrepreneurship, Project Management and IPR:** the student will participate in courses related to entrepreneurship motivation, project management and Intellectual Property Rights.

OTHER INFORMATION OF INTEREST :

EXTERNAL COLLABORATORS

Apart from the respective groups of the co-directors, **for the development of this thesis we will count with the support of external collaborators**, with which co-directors keep an active relationship. These collaborators are well-known experts in the fields addressed in this thesis. They are listed below:

- **José Luis Verdegay:** Professor at the University of Granada (Spain) and Dr. Masegosa's thesis co-director. **Google Scholar h-index: 45**
  Google Scholar profile: https://scholar.google.es/citations?user=7AjYKlcAAAAJ

- **David A. Pelta:** Lecturer at the University of Granada and Dr. Masegosa's thesis co-director. **Google Scholar h-index: 19**
  Google Scholar profile: https://scholar.google.es/citations?user=wrFCypcAAAAJ

- **Francisco Herrera:** Professor at the University of Granada, ISI Highly Cited Researcher and Dr. Triguero's thesis co-director. **Google Scholar h-index: 112**
  Google Scholar profile: https://scholar.google.es/citations?user=HULIk-QAAAAJ

- **Salvador García:** Lecturer at the University of Granada, ISI Highly Cited Researcher and Dr. Triguero's thesis co-director. **Google Scholar h-index: 32**
  Google Scholar profile: https://scholar.google.es/citations?user=vIC06a0AAAAJ

HOST RESEARCH GROUP: Deustotech - Mobility unit

DeustoTech-Mobility unit is committed to respond to the current and future needs of society and industry in the field of mobility and transport by means of technology and artificial intelligence applications. Nowadays, the Unit belongs to **3 European collaboration networks** and participates in **4 European Projects**, as well as is currently developing several consortiums and partnerships for future projects. Actually, **5 international students** are hosted by the Unit, as well as **more than 10 international fellowships have been carried out by members of the Unit**. **More than 70 projects** have been deployed or are under deployment by the unit, as well as **more than 150 scientific contributions** have been done to journal and international conferences.