Car Telematics Big Data Analytics for Insurance & Innovative Mobility Services

Leonardo Longhi, Sistematica S.p.A. , Via G. Peroni 400/402, Rome, Italy

*Mirco Nanni**, ISTI-CNR, Via G. Moruzzi 1, Pisa, Italy – ORCID ID: 0000-0003-3534-4332

*corresponding author: mirco.nanni@isti.cnr.it, tel.:+39-050- 621-2843

Abstract

Car telematics is a large and growing business sector aiming to collect mobility-related data (mainly private and commercial vehicles) and to develop services of various nature both for individual citizens and other companies. Such services and applications include information systems to support car insurances, info-mobility services, ad hoc studies for planning purposes, etc. In this work we report and discuss some of the key challenges that a car telematics pilot application is facing within the EU project "Track and Know". The paper introduces the overall context, the main business goals identified as potentially beneficial of big data solutions and the type of data sources that such applications can rely on (in particular, those available within the project for experimental studies), then discusses initial results of the solutions developed so far and ongoing lines of research. In particular, the discussion will focus on the most relevant applications identified for the project purposes, namely new services for car insurance, electric vehicles mobility and car- and ride-sharing.

Keywords: mobility, big data analytics, car insurance, mobility services, carpooling

Acknowledgements

The work in this paper is partially funded by the European Community Horizon2020 project n. 780754 "Track and Know".

1. Introduction

Mobility data generation and analysis is at the core of the business of many mobility–related companies, including car insurances and associated technology providers. Indeed, providing fresh and detailed information about the mobility of vehicles and single users can be fundamental in optimizing services. This is the case for car insurances, where a good knowledge of the driving attitude of the customer allows to identify the most appropriate contractual conditions, typically associated with the risk of causing accidents. Indeed, risky customers create risks both for their safety and for the car insurance profit, and the best customers for car insurance providers are indeed the safe ones. For this reason, in the long term the business objectives of the company should include not only identifying the risky subjects, but also providing them useful feedbacks to correct their risky behaviours. Similarly, services aiming at supporting alternative transportation solutions, such as car pooling or electric vehicles, require to know which kinds of mobility needs the user has, and then infer what kind of changes to her daily routines are needed to fit the requirements of the new solution. In case of car pooling, that means aligning with the mobility of other users; in the case of electric vehicles, we have to take into consideration the limited autonomy of current batteries, the relatively low availability of recharging points, and the relatively long recharge times.

In this paper we summarize the objectives and the challenges of a pilot application scenario of the EU project Track and Know in the car telematics sector, mainly addressing services and applications in the three areas mentioned above: car insurance, electric vehicle mobility and shared-mobility. In particular, the main goal of the application is to analyze the big mobility data currently produced by car telematics technology providers for routine

tasks (e.g. providing driving statistics for car insurance companies) and then extract insights that can be useful for advanced services.

The data sources generated by car telematics typically include movement traces of vehicles that mount an ad hoc device. Such device periodically establishes the position of the vehicle through GPS technology and also measures other physical characteristics, such as speed and accelerations. In particular, GPS traces are usually collected at a fixed rate or through fixed rules (for instance combining constraints on time passed and distance traveled since last recorded location), while acceleration data are mostly recorded when specific conditions are met, for instance the overall acceleration exceeds some given threshold. One standard functionality of this kind of devices is to produce an alert in case of suspected crash, detected as very large and sudden accelerations, which are timely sent to a human operator to check whether a real crash happened (for instance by calling the vehicle owner) or it was a false alarm.

These data bases provide a good opportunity for developing mobility data analysis models that try to recognize the risk factors behind vehicle crashes, both for being able to predict them and to provide the users indications of how to reduce their (expected) risk. Also, the analysis of long-term mobility needs of a user can provide objective and detailed information about the impact that a change in mobility modalities could have on her daily needs. In the context of electric vehicles, in particular, battery recharge is currently needed more frequently and takes much longer times than fossil fuel-based cars, therefore the habits and timings established in the user's daily activity might change when switching to electric power. Clearly understanding what kind of changes would take place, how big they are, what portion of the mobility they would affect and what ecological and economical impact they would bring, would provide the user the means for taking an informed decision. Similarly, adopting carpooling as (exclusive or complementary) transportation means would clearly require some efforts and changes in the daily mobility. Carpooling can take place only for those movements that have a match with other users' travels, therefore it would be helpful to measure in a data-driven way what is the "carpoolability" ratio of a specific user's mobility. Also, such matches are never perfect, and require the user to anticipate or delay the trip, as well as to move (typically walking) to meet the travel partner. Finally, a complex daily mobility might require the interaction with several different users (the different drivers that give the user a lift), which might make carpooling overall cumbersome and unsustainable in the long run. Clearly, carpooling has positive effects in economical and ecological terms. All these factors, and possibly others, contribute to define pros and cons of carpooling for the single user, helping her to decide whether to adopt it or not as well as companies and public bodies to evaluate the most likely potential of carpooling on a given geographical area.

This contributions of this work can be summarized in two directions:

- first, a set of interesting application-driven analysis problems are defined, some of them new, some others adapted from existing issues;
- second, a set of preliminary results have been obtained on some of the challenges discussed. While far from definitive, the experiments support our initial ideas, confirming their feasibility and potential, which however will require further investigation to turn they into solid and ready-to-market solutions.

In the next sections we briefly present the application context and questions (Section 2), the data sources such applications are based on (Section 3), the main technical challenges identified (Section 4), and some preliminary results over some of the research directions discussed (Section 5). Finally, some conclusive remarks close the paper.

2. Innovative business objectives for car telematics

The Car telematics core business is to collect data from telematics devices and develop advanced solutions and algorithms for sophisticated data analysis, in order to help insurance companies assessing the insurance risks, provide services for the management of accidents, and to facilitate communication between companies and customers. Furthermore, an increasing number of car telematics companies provide services to car manufacturers, the main activities being the following: developing statistics algorithms on individual driving styles and habits, help the car manufacturer to create custom warranty programs derived from driving behavior and offer personalized services to its own customers.

We divide the discussion into the three main application areas of the demonstrator: car insurance, electric mobility, shared mobility.

2.1. Car insurance

Car insurance is one of the most important application fields of car telematics, and the movement data collected by the latter is typically used to provide several services to end users, such as pay-as-you drive contracts, anti-theft control and prompt emergency rescue in case of accidents.

A fundamental task of car insurance companies is to find the most appropriate policy pricing for a customer, which consists in a trade-off between profit and competitiveness. The most intuitive way to do it is to estimate the customer's risk of having accidents in the near future, since high-risk ones are likely to cause the company a loss (paying the costs of her accidents) while low-risk ones are more likely to provide a plain profit. This business case stems from this idea.

The basic objective is not only to recognize the real risk level of a customer, but also to understand possible causes. Therefore, we aim to two distinct results:

- *Predicting the Customer's risk score:* given a car insurance customer, provide a risk score relative to the near future, e.g. the next year or the next three months. We expect this estimate to be greatly dependent on how the customer drives and the conditions of the surrounding environment (traffic, etc.). The methodologies proposed are based on the computation of individual driving features, describing how much the user drives and how much dynamically. More details and preliminary results are given in Section 5.
- *Inferring risk mitigation strategies:* given a car insurance customer and her risk score, we would like to identify the characteristics of her driving that mostly determine her risk score. From a prescriptive viewpoint, that will provide the customer indications of how to improve her risk score, with benefits for her (in terms of safety and insurance costs) and the insurance company (in terms of costs for accidents). The general approach currently under development will try to query the predictive models adopted, in order to understand which features decided for the prediction (see Section 4 for some more details).

As the raw mobility data collected by car telematics companies is limited to positions and events of the vehicle, with no vision of what happens around it, it is clear that in order to achieve our main goals we need to add some information about the context. Similarly, the raw mobility data describes elementary events (position, acceleration, etc.) whereas any proper modeling requires a higher-level vision of what is happening to the user. Such higher-level ones should provide some clear semantics, e.g. some typical maneuvers that involve sequences of deviations, sudden decelerations, etc. Recognizing and making them explicit is expected to be an important need.

Finally, the data involved in this business case imposes several access restrictions that inhibit the end-user of applications to directly access them. The motivations for such restrictions range from individual privacy to competitive advantage of the data provider. Therefore, in order to make the solutions developed practically applicable in an industrial scenario the following important requirement emerges: the data processing that starts from the raw data and terminates with the final results must work essentially unmanned, i.e. without the user interacting or accessing anything but highly aggregated data, e.g. the final risk scores and associated mitigation strategies.

2.2. Electric mobility

While the EVs industry and their adoption is expanding in most EU countries, the switch from fossil fuel to EVs still suffers from a lack of a clear understanding of the pros, cons and habit changes that each user is going to experience. The overall target of this business case is to analyze the mobility of a individual and provide her an objective, data-driven information to detect and quantify possible issues in switching to an fully electrical vehicle. For instance, the limited autonomy of batteries and the current limited availability of recharge stations in some areas might require to heavily change the route of some trips of the user, requiring longer travels and also much longer refill times (battery recharges on average vehicles can take up to some hours, against the few minutes needed for typical gas refills). Such information can help companies and individual users to evaluate the ease of conversion to EV mobility.

In the general context of urban mobility, electric mobility requires the development of new systems that are natively integrated with control, diagnostics and vehicle connectivity systems. With the new systems under

development there will be the possibility for each driver to be able to monitor the performance of the electric vehicle with simple Apps, as well as allowing the use of a lot of information on the status of the vehicle components (e.g. battery charge level, etc.) or to receive alerts in case of interruption of the top-up, unexpected movements, and access to real-time positioning services or sharing of driving data. Furthermore, the growing spread of electric vehicles will also lead to an evolution in the insurance world, since the components that make up electric vehicles, such as batteries, will also be insured. The impact will also be significant in the long-term fleet rental sector, where the transition to electric mobility will favor the spread of new business models, such as pay as you charge, always based on telematics.

The main focus of this demonstrator is on understanding the impact of EV switching on the individual:

- *Estimate Costs/Benefits of EVs for the individual:* given an individual customer with her mobility history, evaluate her costs or savings in terms of money and time in case of switching towards an EV, i.e. provide detailed description of what kind of habit changes, time loss and additional distances traveled the user is expected to incur into, in case an EV is used in her daily mobility. That should take into consideration daily mobility needs, and therefore usual paths, as well as charging point availability (with corresponding detours from the fastest trip) and charging times. The solutions under development in Track & Know will exploit a complete, network-based view of the individual mobility, simulating the battery consumption of the user for her daily trips, and contextualizing possible issues against the part of mobility they affect.

Such general goal will require to understand the mobility needs of the users both at the individual and at the collective level, identifying the most frequent areas of interest or the most frequent or typical routes adopted, as that can help assessing the relevance (weight) of the area or route for the main objective.

A key task involved is to derive the consequences of the limited autonomy and longer recharge times of EVs compared to traditional vehicles, since these two factors might make some fossil fuel-based trip impossible or uncomfortable for an EV. Finding EV-compliant alternative routes for the travels of a user, possibly differentiating among trips of different nature (systematic vs. occasional, long vs. short, easy to substitute with public transport vs. others), and measuring their efficiency constitutes a starting point for the main objective.

2.3. Shared mobility

Sharing vehicles (either through car-sharing or by ride-sharing, i.e. using the user's private vehicle) is a basic measure to reduce traffic congestions, the derived environment footprint, and the personal costs (fuel consumption as well as maintenance costs).

Car-sharing is increasingly becoming a popular, flexible and affordable mobility solution that grows progressively in the metropolis all over the world. It is a simple, ecologically sustainable and alternative paradigm of mobility, especially from an environmental point of view, because it decreases the mean insurance (pay-as-you-drive pricing will decrease, since the personal vehicle is used much less) and maintenance costs of the car (less kms driven mean less vehicle wear), especially in congested urban centers.

The latest research in the area of Shared Mobility predicts that the global carpooling market will grow at a compound annual growth rate (CAGR) of 8% from around 22 million users in 2017 to 47 million in 2025, with more than 500,000 vehicles by 2025 (Frost and Sullivan 2016). For this reason, car telematics companies are venturing into providing technological solutions for Shared Mobility that also include advanced fleet management and insurance telematics for operators in the Mobility sector and for rental companies.

Despite the clear difference between the two cases, EVs mobility and vehicle/ride-sharing have several common points. In particular, both of them would greatly benefit from a clearer understanding of the pros, cons and habit changes that each user is going to experience when she joins it. The overall target of this business case is to provide objective, data-driven means to measure such aspects and let service providers and individual users to evaluate the ease of adoption of car-pooling and/or car-sharing.

We remark that studying car-pooling and studying car-sharing are rather distinct problems, yet they share a large part of concepts and basic tasks, and therefore they are discussed here as a unique subject.
Our main focus here is the following:

- *Estimate Costs/Benefits of car/ride-sharing for the individual*: given an individual customer with her mobility history, evaluate her costs or savings in terms of money and time in case of adoption of car/ride-sharing. That should take into consideration daily mobility needs, and therefore the importance that each trip has in the overall mobility demand of the individual. The solution under development in Track & Know exploit a network-based view of the individual mobility, which makes it possible to find travel partners that not only can share a trip, but whose overall mobility matches the user's one, making it easier to organize the daily mobility.

Synchronizing with other users (to travel together in the ride-sharing case, or to take the shared car in the other case) usually affects the efficiency of the travel in terms of time delays and slight changes in the itinerary. Finding car/ride-sharing alternative routes for the travels of a user (maybe focusing on the most relevant ones) and measuring their efficiency constitutes a starting point for the main objective.

3. Data collection and preprocessing for Car Telematics applications

The main information sources involved in this application context are related to the mobility of individuals (in the specific case, car insurance customers). In particular, all the problems and solutions discussed in this paper are based on the following types of information, for a large set of private vehicles:

- Positions: a list of timestamped WGS84 GPS position (latitude and longitude) related to anonymized vehicles (via anonymous IDs) with an additional labelling about the vehicle travelling status, the satellite connection status and the cardinal orientation of the car. This data is collected at an average rate of one position every 1.5 minutes, though there are some exceptions.
- Events: position data (as above) enriched with threshold-base labels describing motion events occurring in a given times stamp, such as harsh acceleration, harsh braking and (possibly multiple) harsh cornering, with additional accelerometer metrics related to each event position. These data are collected whenever the accelerometer detects an acceleration exceeding predefined parameters (not disclosed to the project).
- Crashes: position data (as above) related to crash events with additional accelerometer metrics (tri-axial average and tri-axial maximum accelerations). This dataset contains all machine-detected crash conditions, basically meaning violent decelerations, including false positives. The records report the result of a manual validation performed by a human operator, therefore distinguishing the true positives from the negative ones.
- Car models: a list of registries about car age, brand and model, related to the anonymous vehicle IDs.

In particular, the data sample involved in the Track & Know project was provided by the OctoTelematics company (www.octotelematics.com), currently the largest player in the global market, and covers three geographical areas (see Figure 1), representing three very different and important situations to be considered in the analyses and services:

- A very large city (London, UK)
- A moderately large city (Rome, Italy)
- A whole region, composed of variable-size cities (Tuscany, Italy)

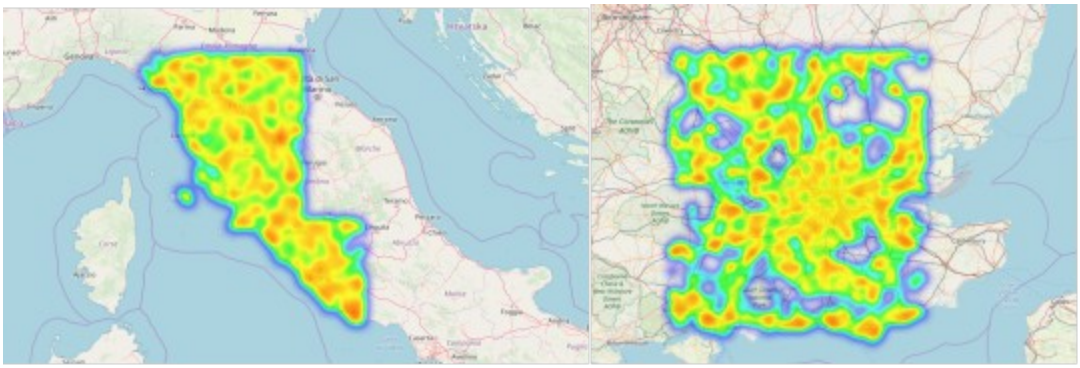


Fig. 1 Geographical areas covered by the data employed in the demonstrator

Also, the data provider has different penetration indexes in the two countries involved (the difference is an order of magnitude), thus providing a natural testbed for analysis tools over heterogeneous data richness, including transfer learning issues.

The large-scale collection of mobility data inevitably brings several quality issues due to a number of causes, yielding either imperfect records (for instance, due to GPS error or incorrect device configuration) or missing ones (for instance, lost data packages). Trying to mitigate such issues requires an ad hoc approach that studies the characteristics of the data sample at hand. In particular, several analyses require the reconstruction of trips out of raw GPS points recorded for each device: determining the start and end of each trip in a precise way requires specific heuristics (e.g. Mousavi, S., Harwood, A., Karunasekera, S. et al. 2017) and, in particular, in our scenario we adopted a spatio-temporal criterion, as described in Section 5; reconstructing the detailed path (which roads were traversed) might require map matching and similar solutions; also, singling out noise and errors is important for obtaining good results, yet, while it is relatively easy to identify large anomalies (which we implemented in our preprocessing steps simply removing points very far from the others), detecting those of moderate size in the data (e.g. a distortion large enough to move a point over the wrong road segment, yet too small to be spotted by visual inspection) can be very challenging.

Finally, several applications require to associate some semantics to the raw data. That is currently realized in the project by simply joining external information, for instance by attaching to each GPS location the weather conditions, local traffic and points-of-interest around it. An alternative, more sophisticate approach consists in inferring such semantics from the available data; for instance, it is currently under study the identification of recurrent trips or the spatial aggregation of driving events aimed to identify areas were some specific behaviours are more frequent, e.g. bad road conditions leading to frequent sudden decelerations.

4. Technical challenges and related works

The business cases described in the previous section present several challenges from the technical viewpoint, since they mostly require a deep understanding of human mobility starting from raw data lacking any detailed semantics. In this section we discuss some of the most important ones, linking them to existing literature and highlighting the specificities of our context.

4.1. Individual-centered mobility modeling

A specific type of semantics is related to the meaning that the different parts of the mobility have for the individual: recurrent vs. systematic trips, frequent locations vs. single visit ones, transit locations vs. long stays, etc. To infer this type of information we need to model the mobility of the individual as a whole, creating a single, complete picture of it. This process is currently ongoing exploiting Individual Mobility Networks (Rinzivillo et al. 2014), a network-based representation that integrates important locations, movements and their temporal dimension in a succinct way. Such model allows several different types of inference (detecting the purpose of the trip,

simulating realistic mobility agendas, etc.), in contrast to others that are tailored around a specific objectives, e.g. predicting next location (e.g. Amirat, H., Lagraa, N., Fournier-Viger, P. et al. 2019). Integrating as much information as possible in a single formalism and inferring from it mobility indicators useful for the predictive/prescriptive purposes of the demonstrator are among the key challenges.

4.2. Prediction of (crash) risk probability

Risk in this context means probability of accidents, which are (in statistical terms) rare events. That, together with the lack of a clear set of predictive indicators to adopt, make the risk prediction a difficult task.

The existing literature addresses the problem from various perspectives. A large body of works focus on real-time prediction of individual crashes, i.e. try to identify the events that lead to a crash in next few seconds, thus providing feedbacks to the user as she drives, e.g. Wang, Xu and Gong (2010). Similarly, though following completely different directions, Yutao B et al. (2017) try to related crashes to both behavioural characteristics and physiologic parameters. Other approaches work on identifying areas that show characteristics usually associated with accidents, such as increased traffic density, adverse weather conditions, etc., e.g. Lee, Hellinga and Saccomanno (2003) and Mannering and Bhat (2014). While extremely useful, such approaches result to be not applicable to fields like car insurance, where we are interested in creating a general risk profile of the user, thus implicitly involving the prediction of her crash risk in the long run, such as few months in the future. Only few, early works are available on this direction, e.g. Wang et al. (2017), limited to simplistic approaches.

The approach under development will take into consideration several aspects, ranging from the driving behaviour of the user to the types of environment she usually traverses – the latter includes both static information, such as road categories, and dynamic ones, such as weather during driving time.

4.3. From prediction to prescription

Achieving a good prediction accuracy often conflicts with the understandability of the predictive model. It is well known that in difficult settings very complex models (deep learning, large random forests, etc.) can achieve far better performances than simpler ones (decision trees, Bayesian classifiers, etc.); yet, the former are usually not human understandable. One of our main objectives is not only to provide good predictors for the car crash application, but also extracting risk mitigation guidelines for the user (the driver), which means we are interested in understanding which factors made a driver a risky one, in order to propose changes in her behaviour that can reduce the risk. While that makes simpler models more appealing, the project will explore also methodologies coming from the “explainable AI” community (e.g. Guidotti et al. 2018), aimed to extract from a black-box model an explanation for each prediction obtained. Current work within the project is addressing the problem exploring approaches based on adversarial learning (Kurakin A, Goodfellow IJ and Bengio S, 2017), which traditionally tackles similar problems yet with very different purposes, and counter-factual analysis (see e.g. Poyiadzi R et al. 2019).

4.4. Models Transferability

The various types of mobility models involved in this demonstrator are expected to be highly dependent on the specific geographical area under study. For instance, it has been empirically verified that the trip purpose prediction models proposed by Rinzivillo et al. (2014) work very well in the areas where they were extracted, their performances degrade dramatically if applied to areas with different characteristics. At the same time, not all areas of interest for the demonstrator are equally well covered by data, due to the non-homogeneous penetration of tracking devices, making it difficult to build different models for different areas. For instance, the penetration of GPS vehicle trackers in UK is an order of magnitude lower than Italy, and other countries where this market just started show even lower values. All this calls for methodologies that make it possible to adapt models built in data-rich areas to less rich ones, basically a geographical instance of the general transfer learning problem (Pan and Yang 2009).

4.5. Defining proper notions of electrificability and shareability of individual mobility

Measuring how much the mobility of an individual is compatible with alternative transport modalities – in our case, EVs and car sharing/pooling, both with their own constraints – is a not well defined problem. Existing work measured the ratio of trips that are perfectly compatible with them (Guidotti et al. 2017, Janssens et al. 2012), or simply compare general mobility demand (based on trip length distribution and other overall descriptors, for instance, as in Donati A et al. 2015) but without a more realistic evaluation of the effort required on behalf of the user to adapt her whole mobility. In the case of EVs, that means changing times and routes to intercept charging stations when needed. Moreover, mobility optimization might intersect energy distribution issues, including the balance of energy consumption on the grid or how to use vehicles as potential distribution means, as studied in Neaimeh et al. (2015); for car sharing/pooling, it means to change times of travels, or even reschedule part of them. Providing such definitions and the tools for computing their values is another challenge the project (and this demonstrator in particular) is going to pursue.

5. Preliminary results and insights on the data

This section summarizes some of the first insights and preliminary results obtained over the datasets adopted in the demonstrator, and focused on the main demonstrator objectives and most promising analytical tools.

All experiments are based on a trajectory extraction process that scans the raw GPS traces of an individual in chronological order, filters out noisy points (here defined as those whose distance from the previous point would imply an average speed above 250 km/h), and identify stops (here defined as moments where the vehicle moved less than 50 meters in the last 20 minutes). A trajectory is then defined by the points between two consecutive stops.

5.1. Mobility-based Characterization of Geographical Areas

Most analyses and models extracted from data are highly dependent on the characteristics of the territory under study. In particular, it is known that mobility models extracted in one region might not work well in other ones, thus raising an issue of transferring models across different areas. In this direction, the technical activities of the Track & Know project are addressing the problem of characterizing different areas based on a wide variety of indicators, with the aim of better assessing the similarity of different geographical areas (the idea being that models are more easily transferrable between similar areas) and possibly devise mechanisms to adapt models across areas with different characteristics.

The initial exploration on this line considered the following families of mobility-based city indicators:

- Spatial Concentration of population: various measures of concentration are computed over each city, including spatial entropy and Moran's I, based on a fixed tessellation of the territory.
- Traffic flows distribution: starting from the traffic network among the sub-areas of a city, various indexes are computed, such as the modularity index (Newman 2006), as well as the fitness of such traffic distribution with standard mobility models like the gravitational model (IZA World of Labor 2016).
- Distribution of IMNs properties: for each individual estimated to be resident in the city, we build his mobility network (Rinzivillo et al. 2014) and analyze its network features, such as number of nodes, etc.
- Road network and traffic concentration: the static structure of roads in the city is analysed, by computing for instance their spatial concentration, and by joining them with real mobility data we measure how much the traffic is concentrated in a few km of roads.

Examples of the above mentioned measures are shown in the following figure, plotting their spatial distribution over the Tuscany region. It is clearly visible that most indicators have a rather high heterogeneity over the territory, meaning that each city shows some difference from others, including close ones. At the same time, each indicator is significantly different from the others, thus bringing potentially useful and non-redundant information.

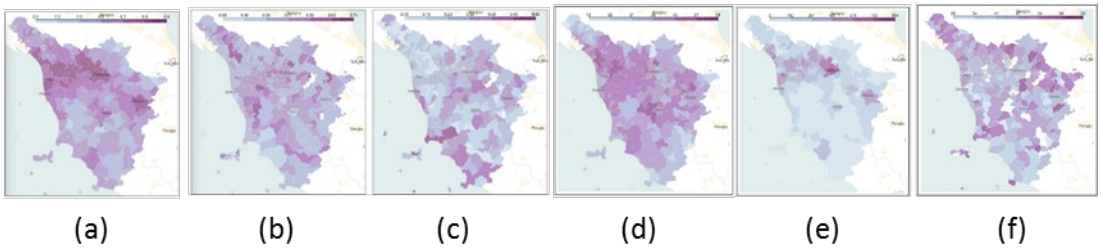


Fig. 2 Spatial distribution of sample city indicators over the Tuscany region: (a) Population entropy, (b) Modularity, (c) Fit to gravitation model, (d) N. of nodes in IMNs, (e) Roads concentration, (f) Traffic concentration

Discussion of results: the experimental results obtained confirmed our working assumption about the possibility of identifying local, discriminating properties of a territory looking at its mobility in conjunction with its geography. The usability of such features has been tested on some specific analysis task (Section 5.3), yet it is still to understand how well they capture the phenomena they describe, and whether they are correlated with other contextual features.

5.2. Individual Mobility Networks

Based on the paradigm introduced by Rinzivillo et al (2014), the mobility of an individual can be summarized by a graph representing the locations visited by the user (inferred from the single start- and end-points of each trip performed) and the transitions between locations, together with spatio-temporal distributions associated to each location and transition. IMNs are a basic tool to analyze the population of an area through the characteristics of the individual that live there. First explorations show that the differences across different areas are not easy to spot through direct visual inspection, as shown in the following figure.

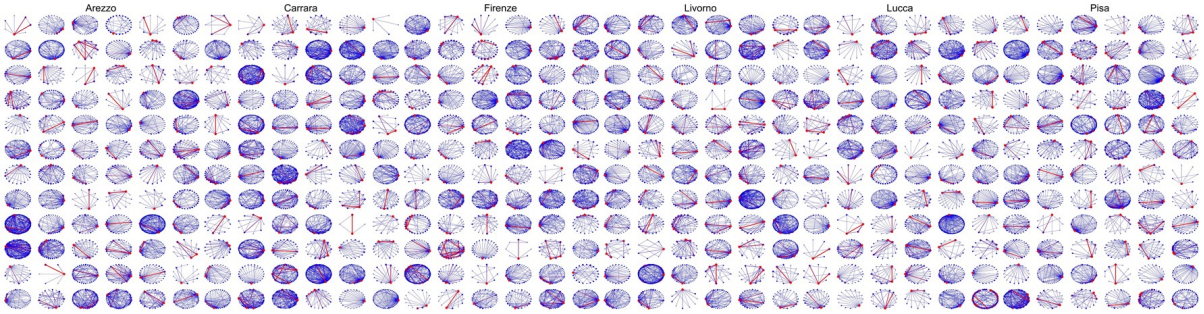


Fig. 3 Sample IMNs for 6 different cities in Tuscany; apparently, no clear visual feature characterize cities

Therefore, new ways of representing, aggregating and visualizing IMNs are under study, to enable a more effective comparative analysis of different territories.
In addition to that, human mobility is a dynamic phenomenon that can change significantly in time, and therefore IMNs can represent the gradual evolution of users' changing mobility needs. The following figure shows an example where the IMN of a user has been computed over two months (left) and then recomputed over the following two months (right).

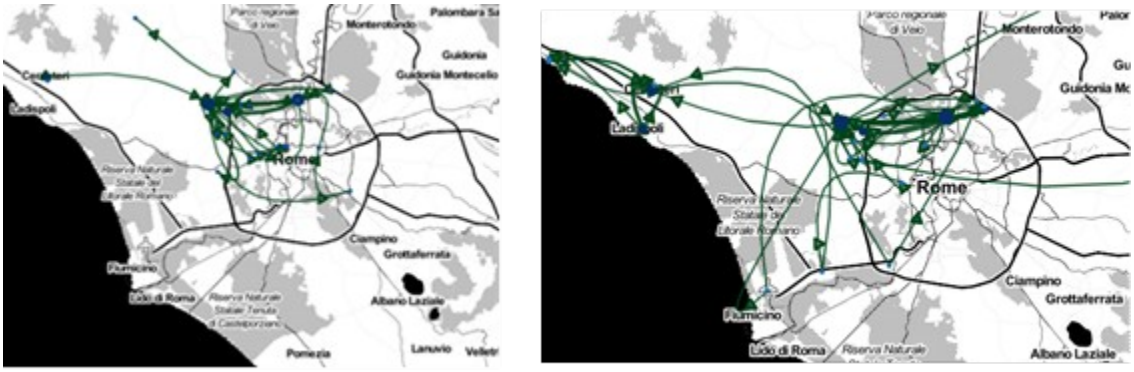


Fig. 4 Temporal evolution of a IMN computed over 2–months periods; changes in mobility are clearly visible

In can be easily seen that while the core parts of the user's mobility are preserved (North of the city), its spatial range extended significantly over new areas. Also, the frequency of visits of the new areas (represented by the size of the corresponding nodes) suggest they became part of the user's routines. How to integrate such evolution patterns in a comprehensive model of human mobility is a challenging question that is currently under study.

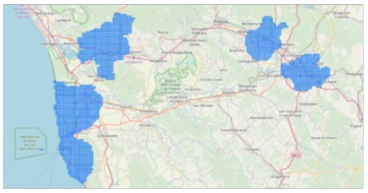
Discussion of results: the preliminary results confirm the usability of IMNs for summarizing individual mobility, yet also showing that the level of detail of the identified locations can be sometimes too fine (each user has several tiny locations that are difficult to analyze), and that the time component in long-duration data must be considered in a more comprehensive way, for instance developing dynamics-aware models.

5.3. Geographical Transfer of Mobility Models

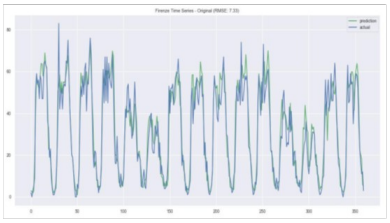
As previously discussed, the data availability in the mobility domain is often heterogeneous, allowing to build strong models (for instance, predictive ones) on some geographical areas where rich data are accessible, but not on other, less rich areas. It is well known that any (non-trivial) mobility model is tightly linked to the area it describes, therefore we expect that a good model built on a (data-rich) area does not work equally well on different regions.

There are two main ways to tackle the problem. One consists in developing a strategy that takes a strong model built on a specific area A and adapts it to work on a different area B by exploiting the (relatively little) information available over the latter. The already cited approaches in Rinzivillo et al (2014) represent examples of such line of work. Another, simpler way consists in recognizing which are the areas where the model developed on area A is likely to perform well. This clearly requires to study the features of the areas that make them somehow compatible, i.e. they apparently obey to the same kind of rules. In the following we briefly report some results obtained on this second direction, where the city descriptors introduced in a previous section (Mobility-based Characterization of Geographical Areas) have been deployed to group cities into clusters.

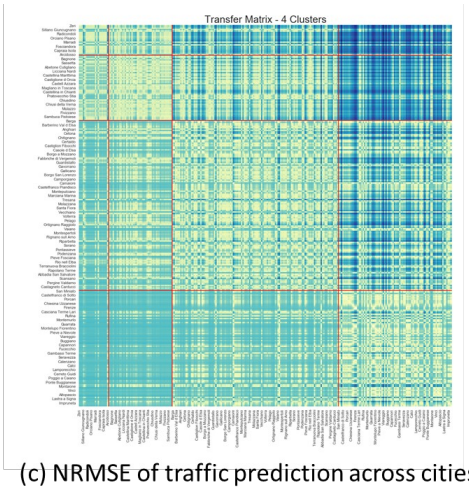
First, a simple prediction problem is defined: predicting the traffic of the next hour in key areas of a city. In particular, each of the 270+ municipalities of Tuscany, Italy was divided in a regular grid and 10 representative cells were selected, 5 among the top 10% traffic and 5 within the 80-90% percentile of traffic. For each city, then, a time series representing the aggregate hourly traffic volume of such cells is obtained, and the prediction task is to predict the next value based on previous ones. The prediction model adopted is a standard XGBoost regressor (Chen and Guestrin 2016). The next figure shows four sample cities analyzed (a) and a sample hourly time series (b).



(a) Sample cities and grids



(b) Sample traffic time series



(c) NRMSE of traffic prediction across cities

Fig. 5 sample cities analyzed (a), sample hourly time series and (c) normalized root mean square error (NRMSE)

The matrix in Figure (c) shows, for each pair (A,B) of cities, the normalized RMSE (root mean square error) of predictions obtained on B by using the model learnt on city A. The cities have been clustered through a hierarchical agglomerative method based on the city features already introduced in Section 5.1, which yielded 4 clusters. The matrix described above has the rows and columns sorted according to the cluster each city belongs to, resulting in a block matrix, blocks being delimited by red lines. What we can see, is that blocks on the diagonal, corresponding to cities in the same cluster, exhibit a brighter color than others, which means that the NRMSE error tends to be smaller. This provides a first evidence that the cities that look similar based on the features studied above are also homogeneous in terms of rules that drive the evolution of traffic volumes, and therefore the prediction model is more easily transferrable among them.

Discussion of results: the results shown above are obviously just a first step towards the overall objective. Indeed, the prediction problem and the model adopted were rather simple, whereas in real applications, including those considered in this paper, both problem and model are expected to be much more complex and challenging. Moreover, the approach followed here, i.e. recognizing pairs of model-compliant cities, is effective only if the data-rich cities available cover most of the city types (the clusters in our experiment) we expect to meet, since in that case each city can be served with a model built from a data-rich city of the same type. Data-poor cities of different types would be not associated with any model. The more general solution consists in defining adaptation strategies (possibly based on the same city features considered here) that allow to customize a model to the specific city we need to apply to.

5.4. Crash Prediction

Predicting the crash risk of a user is a difficult task, since it is in general affected not only by how the user drives, but also by external factors, including other drivers. As already discussed in Section 4, most works in literature focus on real-time prediction of individual crashes, or on the identification of personal or contextual factors that relate to crashes. In the car insurance domain we are interested in creating a user's risk profile related to long periods of time, such as months in the future.

In our this preliminary exploration of the problem, we focused on such long-term prediction of crash risk, and we measure what kind of performances we can expect to reach with simple users' features. In particular, experiments consist in characterizing each user by his mobility data in a time window of three months, and try to predict the presence of crashes in the next month. The experiments include only users that have a significant mobility (here defined as those making at least 10 trips in the period under observation), since inactive vehicles are not interesting

for our purposes – their crash risk is virtually zero. No data balancing or other particular filtering was performed, yet for practical reasons the experiments focused on a time period where the density of crashes was the highest.

The features adopted fall in the following three categories:

- Travel features: length and duration of trips, also split into periods of the day or of the week
- Events features: frequency and intensity of driving events, i.e. accelerations and decelerations, divided by event type and temporal intervals
- Car brand and model

The prediction was performed with various methods, including Random Forests, Support Vector Machines and Neural Networks. RFs yielded the best and more stable results, shown in the following figure. The table also divides performances over different subsets of features (traj = travel only, evnt = events only, evnt = both, all = include also brand and model). The results were computed over a sample of data, covering vehicles in Rome and London, and the corresponding model parameters were selected by grid search optimization.

clf	features	f1-score	precision	recall	test_accuracy	train_accuracy
RF	all	0.659024	0.709972	0.655093	0.712644	0.815271
RF	evnt	0.641626	0.734188	0.663095	0.672414	0.757389
RF	traj	0.636443	0.723219	0.655688	0.669540	0.748768
RF	trev	0.603573	0.651994	0.609672	0.658046	0.795567

Fig. 6 Performances of Random Forest models on various sets of features

We can see that using all feature types the overall performances (F1 score) is maximized, and therefore all features appear to bring some improvement. We notice that the problem is imbalanced (around 1 crash every 5 users), therefore a significant recall is as valuable as a high accuracy.

Discussion of results: the results obtained show that the problem can be approached with the methods discussed above, although the results still call for technical improvements. In particular, current ongoing work is integrating other, more sophisticated features that take advantage of the IMNs of the users and of contextual information.

6. Conclusions and future works

This paper presented a set of challenges in the car telematics domain, that correspond to a pilot application of the Track and Know EU project, focusing in particular on telematics car insurance and mobility services. The technical challenges to transform the raw mobility data collected by the telematics companies into insights and valuable services are numerous and require improvements of current research state-of-art. Preliminary results show promising signals of meaningful solutions for the identified problems.

The ongoing work is trying to purse several of the issues mentioned in the paper: developing more sophisticated individual mobility models, that might extend existing Individual Mobility Networks; developing strategies to adapt one model built in a geographical area to work well on a different one; developing a set of sophisticated mobility descriptors to better identify crash risks in the long term, including relations with the geo-spatial context, weather conditions, changes in the driving habits (e.g. through the analysis of the user's IMN changes), etc.; defining satisfactory indicators to measure the compatibility of users with shared mobility or electric vehicles, as well as developing processes to accurately estimate them.

References

Amirat, H., Lagraa, N., Fournier-Viger, P. et al. (2019) NextRoute: a lossless model for accurate mobility prediction. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-019-01327-w

Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. 22nd SIGKDD Conference on Knowledge Discovery and Data Mining.

Donati A et al. (2015). Individual mobility: From conventional to electric cars. EUR - Scientific and Technical Research Reports. DOI: 10.2790/405373.

Frost and Sullivan (2016) Future of Carsharing Market to 2025. Technology Advancements, Market Consolidation and Government Initiatives to Influence Market Growth Over the Next Decade. https://store.frost.com/future-of-carsharing-market-to-2025.html

Guidotti R et al (2018). A survey of methods for explaining black box models. ACM Computing Surveys (CSUR) Surveys, Volume 51 Issue 5, Article No. 93.

Guidotti R, Nanni M, Rinzivillo S, Pedreschi D, Giannotti F (2017) Never Drive Alone: Boosting carpooling with network analysis. Information Systems journal (IS), Volume 64, Pages 237-257.

IZA World of Labor (2016). Gravity models: A tool for migration analysis. https://wol.iza.org/articles/gravity-models-tool-for-migration-analysis

Janssens D, Giannotti F, Nanni M, Pedreschi D, Rinzivillo S (2012) Data Science for Simulating the Era of Electric Vehicles. Kunstliche Intelligenz (KI), 26(3): 275-278.

Kurakin A, Goodfellow IJ and Bengio S (2017). Adversarial Machine Learning at Scale. Google AI, https://arxiv.org/abs/1611.01236.

Lee C, Hellinga B, Saccomanno F (2003) Real-Time Crash Prediction Model for the Application to Crash Prevention in Freeway Traffic. Transportation Research Board, 82nd Annual Meeting, January 12-16.

Mannering FL, Bhat CR (2014) Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research 1, pages 1–22.

Mousavi, S., Harwood, A., Karunasekera, S. et al. (2017) Geometry of interest (GOI): spatio-temporal destination extraction and partitioning in GPS trajectory data. J Ambient Intell Human Comput 8: 419. https://doi.org/10.1007/s12652-016-0400-5

Neaimeh M et al (2015). A probabilistic approach to combining smart meter and electric vehicle charging data to investigate distribution network impacts. Applied Energy, dx.doi.org/10.1016/j.apenergy.2015.01.144.

Newman MEJ (2006) Modularity and community structure in networks. PNAS, vol. 103, nº 23, pp. 8577–8582.

Rinzivillo R, Gabrielli L, Nanni M, Pappalardo L, Giannotti F, Pedreschi D (2014) The Purpose of Motion: Learning Activities from Individual Mobility Networks. Int. Conf. on Data Science and Advanced Analytics (DSAA14).

Pan SJ, Yang Q (2009) A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering.

Poyiadzi R et al. (2019). FACE: Feasible and Actionable Counterfactual Explanations. Arxiv: https://arxiv.org/abs/1909.09369.

Wang J, Xu W, Gong Y (2010) Real-time driving danger-level prediction. Engineering Applications of Artificial Intelligence. Volume 23, Issue 8, Pages 1247-1254.

Wang Y et al. (2017) Machine Learning Methods for Driving Risk Prediction. In ACM EM-GIS'17.

Yutao B et al. (2017). Crash prediction with behavioral and physiological features for advanced vehicle collision avoidance system. Transportation Research Part C: Emerging Technologies. 74. 22-33.