Alma Mater Studiorum - Università degli Studi di Bologna

DIPARTIMENTO DI FISICA E ASTRONOMIA

Dottorato di Ricerca in Fisica - XXV Ciclo

STATISTICAL PHYSICS AND MODELING OF HUMAN MOBILITY

Riccardo Gallotti

PhD Thesis

Coordinatore Dottorato:
Prof. Fabio Ortolani

Relatori:
Prof. Sandro Rambaldi
Prof. Armando Bazzani

Settore Concorsuale 02/B3 - Settore Scientifico Disciplinare FIS/07

Anno 2013

Abstract

In this thesis, we extend some ideas of statistical physics to describe the properties of human mobility. By using a database containing GPS measures of individual paths (position, velocity and covered space at a spatial scale of $\simeq 2$ Km or a time scale of 30 sec), which includes the 2% of the private vehicles in Italy, we succeed in determining some statistical empirical laws pointing out "universal" characteristics of human mobility. Developing simple stochastic models suggesting possible explanations of the empirical observations, we are able to indicate what are the key quantities and cognitive features that are ruling individuals' mobility.

To understand the features of individual dynamics, we have studied different aspects of urban mobility from a physical point of view. We discuss the implications of the Benford's law emerging from the distribution of times elapsed between successive trips. We observe how the daily travel-time budget is related with many aspects of the urban environment, and describe how the daily mobility budget is then spent. We link the scaling properties of individual mobility networks to the inhomogeneous average durations of the activities that are performed, and those of the networks describing people's common use of space with the fractional dimension of the urban territory. We study entropy measures of individual mobility patterns, showing that they carry almost the same information of the related mobility networks, but are also influenced by a hierarchy among the activities performed. We discover that Wardrop's principles are violated as drivers have only incomplete information on traffic state and therefore rely on knowledge on the average travel-times. We propose an assimilation model to solve the intrinsic scattering of GPS data on the street network, permitting the real-time reconstruction of traffic state at a urban scale.

Contents

Chapter 1

Introduction

The characteristics of people movements are very important pieces of information upon which many models and studies depend. Individual mobility is coupled with the spatial distribution of activities in a city, a fundamental problem in geography and spatial economics[1]. At the same time, human mobility at all scales is a key ingredient for epidemics spreading models, as diseases are transmitted at close proximity and diffuse because persons travel and interact[2]. The need for transportation is realized by participating to traffic flows over the street network, in most cases using vehicles. When the flow of vehicles overcomes the capacity of a street, a transition to a congested state takes place[3]. These transitions have direct costs in terms of time and monetary loss. Each year traffic delays in the US are said to cost nearly $100 billion, and waste around 10 billion liters of fuel, while drivers in Los Angeles can expect to spend a total of 56 hours sitting in jams[4]. Indeed, the importance of handling traffic was already manifest at the time of ancient Rome: in order to reduce the frequent congestion that afflicted the narrow and crowded roads of the city, Julius Caesar forbade wagon traffic from dawn until dusk, ratifying the first traffic management law in the history[5]. Therefore, the study of the statistics of humans movements and their interactions is fundamental for urban and transportation planning, epidemics and traffic congestion containment, and more generally for the design of smarter cities[6].

Even if we have already made many progresses in the understanding of the microscopical behavior of vehicular and pedestrian flows[3], when we wide the scope to the general picture of human mobility, integrating the complex features and heterogeneities of real-world systems, we are challenged by what in principle is a daunting many-body problem that can be included in the class of complex techno-social systems[7]. Unlike many physics problems, in this case the nature of the

complex of factors determining individual mobility choices and the way people interact with their surroundings cannot be completely known. This kind of systems can hardly be completely understood from a particular disciplinary perspective, and probably the very finding of a single, consistent, complete and correct model for them is unattainable[8].

The objective of this thesis is to participate to the effort of understanding human mobility from a statistical physics viewpoint, and at the same time to study human mobility as a paradigmatic example of a complex statistical cognitive particles system. If we could assume that statistical laws are governing individual dynamics, we could make predictions on system evolution averaging over the evolutions of an ensemble of possible states. This ensemble is characterized by macroscopic observables that give information on the global state of the system. Knowing the nature of this observables permit us to use them as control variables, which are essential to understand the macroscopic behavior of the system, its transient and critical states and phase transitions. Although, the microscopical laws governing individual dynamics are not known, and is therefore necessary to start with the statistical analysis of individual mobility and understand in what measure our ignorance of the micro-dynamical details can be overcome by statistical physics' methods and, alternatively, in which situations the complexity of the system makes those details essential. Studies in this sense have become possible in recent years thanks to the availability of vast amounts of data produced by the Information and Communication Technologies. Mobile phones, GPS devices and geo-referenced social networking are a continuous source of data on peoples whereabouts. This data abundance makes even more important to bring at light where the important information lies, and the specific development of aggregation, filtering and analysis methods is needed. In fact, the potential power of this huge data flow is limited by our ability of extracting the information really needed to create models which can be used to anticipate trends, evaluate risks, and eventually manage future events[9]. Moreover, when studying a cognitive system, we cannot limit our modeling to the description of how from micro (the individual dynamics) we obtain the macro (the population dynamics). The awareness of all individual of the conditions at macro level and the ability of changing strategies implicate also an adaptive feedback in microscopical dynamics. The way people choose among different strategies, the quantities considered in these decisions and the effective understanding of the real state of the system and of its evolution have to be taken into account in modeling population dynamics.

Our modeling approach is data driven. From the analysis of a large GPS database of single-vehicle mobility we deduce statistical laws valid over the whole population. Then we design stochastic models that, capturing the fundamental features of individual mobility strategies, are able to reproduce the observed statistical properties. These models, lying conceptually between micro and macro level, permit us to point out what are the key quantities required for the description of human mobility, and the ways these quantities concur into individual decisions. The knowledge of those quantities and their properties allows to a deeper analysis of individual dynamics, of which we are able to isolate and study new aspects subtracting what we are already aware of. On the other hand, quantifying the relationships of those key quantities with the physiological, economical, infrastructural or geographical constraints influencing our mobility, we can reach the goal of characterizing the strategical processes underlying driver's decision dynamics, making one or more steps forward in the developing of a mobility governance framework.

Chapter 2

Mobility Data

FIGURE 2.1: Distribution of driver's presences in our database.

This work takes advantage of a huge database of GPS (Global Positioning System) measurements, describing the motion of private vehicles in Italy. Those data have been collected, primarily for insurance reasons, by a private company (Octo Telematics S.p.A.[10]), who granted us access to part of its database for research purposes. This database refers to roughly 2% of the vehicles registered in Italy. The big advantage in the use of GPS measures for the study of human mobility is the chance of directly following the movements of people. In fact, we can easily and precisely define a trip as the transfer between two places where the engine has been turned off, where other indirect measures of human mobility, such phone calls' location or of accesses to a social network, may be systematically influenced by the complex features of the communication habits. Moreover, this type of data clearly permits to study at the same time the individual mobility in the urban environment and the associated use of the road network.

The installed GPS devices can record and send to the main Octo Telematics Service Center the geographical coordinates, time, instantaneous velocity, distance covered since the last record and quality of the GPS signal at the time of the recording, that can be absent, weak or good. Communications with the Service Center are made via GSM/GPRS network. Therefore, to reduce costs, the records are normally taken only at engine start (starting data), stop (stopping data) and every approximately 2 Km of travel (travel data) and sent in packets of 50. In particular, after 2 Km from the last record, the device will save a new one as soon as it has a sufficiently good GPS signal. A slightly different recording pace is used for travels along highways or ring roads around important cities, where the company provides traffic information in real time. In this case, when a car enters the highway, it passes through a virtual gate and its position is recorded. Starting from that moment the record will be made every 30 seconds and sent every 12 minutes until a second passage through another virtual gate happens exiting the highway. Being this system not perfect, it may happen that a car is recording with a 30 second pace outside highways.

The data suffer from the limited precision of GPS measurements, in particular when the device looses the satellite signal. These problems are especially relevant when the engine is switched on or the vehicle in parked inside a building. When the signal quality is good, the time precision of the recorded data is perfect, whereas the space precision is of the order of 10 m, which is usually sufficient to localize a vehicle on the road. In adverse circumstances, errors can increase up to 30 meters or more. Naturally, errors on both instantaneous velocity and covered space are related to errors on the GPS positioning given. Nevertheless, these last quantities

are calculated with an adequate precision, as they result from an elaboration of GPS data recorded (but not registered) each second.

Due to the Italian law on privacy, we have no direct information on the owners or any specific knowledge about the social composition of the sample. The installation of this GPS system on a vehicle entitles the holder to a discount off the insurance price. This is particular appealing for young people so it can be expected a bias in this sense. Taxi companies or the delivering services use their own GPS systems and they do not contribute to the database, which is mainly set up by private vehicles. There is a small percentage of vehicles used for professional reasons and belonging to private companies, that take advantage of the insurance discounts of collective contracts.

In this Thesis we analyze three datasets containing the records of one month of mobility in different areas and years. The first dataset is about the province of Florence in March 2008, the second about the region Emilia-Romagna [1] in November 2009 and the last dataset contains information about all the mobility in Italy in May 2011. More information about the dimensions of these datasets are found in table 2.1

	Month	Surface	Trips	Cars
Florence	March 2008	3,514 Km^2	1,806,000	32,000
Emilia-Romagna	November 2009	22,451 Km^2	7,157,000	75,000
Italy	May 2011	301,340 Km^2	128,363,000	779,000

TABLE 2.1: Datasets dimensions.

In the following, I will illustrate some of the algorithms that have been developed within the Physics of the City research group for the analysis of these datasets.

2.1 Data Pre-Elaboration

GPS data are originally organized chronologically. In order to consolidate the trip structure, all measures are re-organized with a new ordering given by the vehicle identifier. Then, it is verified if trip sequences (starting datum - travel data - stopping datum) are complete. In this phase, all measures without informative

[1] For Emilia-Romagna the analysis has been made considering only working days.

content, i.e. double records or travel measures taken without satellite signal, are erased. Between measures taken too close one another ($\leq$ 30 m or 20 sec), only one is kept and if an entire trip is, in this sense, too short, it is completely erased.

In cases where the starting and stopping data are taken with low levels of signal, other measurement associated to the same stay are used to have a quality improvement. This happens often for data taken when the engine has started: as the device has been just turned on, this kind of data is often taken with no signal. In all cases, when there is no signal, the device records the last known good quality position as the actual one. In the case of starting data, the last known position should be the place where the engine has been turned off, and thus probably describing to the same location. If the two measured positions are distant less than 100 meters, the quality of the starting point data has been enhanced as it is supported by the precedent stopping point, otherwise we may have a pathological stop, where the continuity of the trajectory is lost. Moreover, for all starting data, both velocity and covered length of the starting point are initialized to a nil value.

Another situation where the continuity of the trajectory is lost is when a signal loss occurs during the travel. This can happen for many technological or environmental reasons. If, as a result of the signal loss, two consecutive GPS data, with good or weak satellite signal, are too far one another (temporally: more than an hour; or spatially: more than 3 Km), the trip is considered interrupted and we declare no knowledge on what has happened in that tract.

Finally, it has been observed it that in many cases reasonably continuous trajectories were interrupted by stops (identified by a stopping and a following starting datum) that were hardly justifiable. In particular, the duration of these stops was too short to be associated to any activity done out from the vehicle. It could be both that the engine was turned off for a little while, voluntarily (i.e. at traffic lights) or not, or a consequence of technical problems with the device. In any case, these short stops have been filtered out: two consecutive trajectories interrupted by a stop shorter than 30 seconds are joint in one longer trajectory. As an exception, the stop is kept if the angle formed by the stopping position and precedent and following points in the trajectory was not far from 180 degrees, as probably the driver is making a round trip with the purpose of giving someone a lift.

In the pre-elaboration phase have been eliminated roughly the 10% of the raw GPS data, while the 17% has been modified.

2.2 Parking Points Clustering

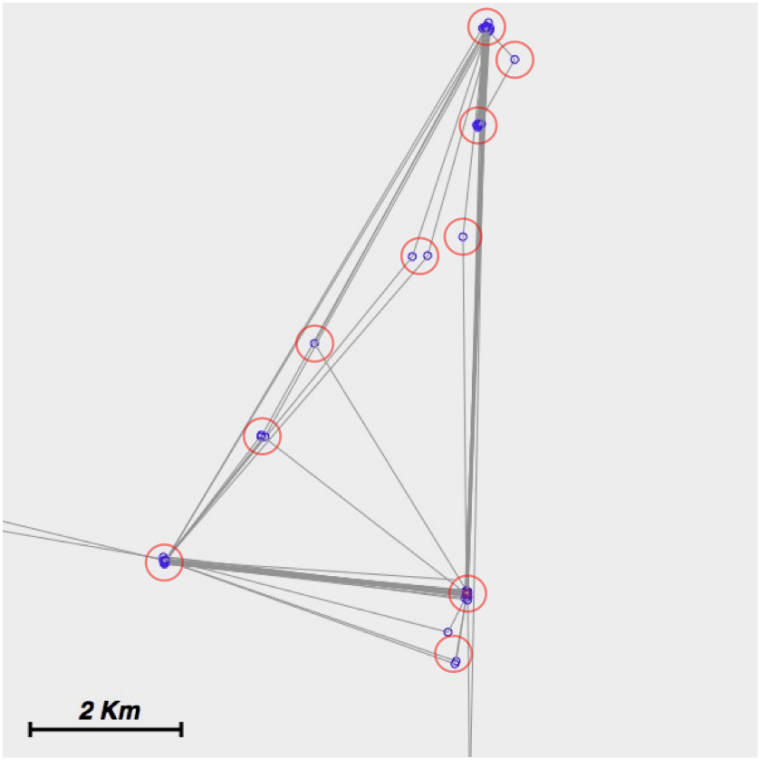


FIGURE 2.2: A result of the clustering algorithm.

One of the objective of our study is the identification of the decisional mechanisms at the root of each individual's mobility planning. Our data provide naturally a good microscopical description of this mobility. However, it is our interest to be able to recognize different locations visited by the single drivers (home, workplace, ...), that we identify with the coordinates where each trip ends, the parking point, and we associate to an activity performed near the parking place. To bi-univocally associate parking coordinates with visited locations is not trivial, since we have a precision that can even resolve two different parking places in the same parking area, and at the same time an high risk of signal loss errors due to technological limits and the use of underground parking places. Hence, from the cloud of all the parking points, activity locations have been identified through a gravitational clustering algorithm. This mechanism is based on the assumption that between the location where an activity is carried out and the chosen parking place there is a maximum acceptable parking distance distance[11]. This distance has been first assumed of 400 m for the Florence dataset and then modified to 500 m for the Emilia-Romagna and Italy dataset. This difference did not appear to

have noticeable consequences on the obtained statistical results. In addition, for Florence and Italy the clustering procedure has been performed individually, i.e. the parking places of different individuals do not interact for the purposes of the clustering. In the used gravitational clustering algorithm, at each visited point is associated an unitary weight. Under the threshold value given by the maximum parking distance, two different locations are joint in another, which weight is the sum of those of the original two and which coordinates are the barycenter. This aggregation takes place iteratively, picking at each time the two closest points in the ensemble, until those are further than the threshold value. As long as we consider separately the different individuals, the algorithm is computationally efficient. In any case it has the advantage of bringing to a one-to-one result, not depending by an arbitrary choice of the first aggregation points, but only on the maximum parking distance.

2.3 Map Matching and Path Reconstruction

FIGURE 2.3: **(Left)** Map Matching and **(Right)** Trajectory Reconstruction.

The information carried by our GPS dataset is naturally bound to the street network, where all vehicular movements take place. Nevertheless, we know only geographical coordinates and velocity of a vehicle, which are not originally related to the position on one particular road. The shift from coordinates to roads represents a precious enrichment of our knowledge on the nature of the trips performed. This knowledge is essential for studies on road traffic, where the street network plays a central role. In this thesis, those analysis are part of chapters 7 and 8.

The passage from the recorded coordinates to trajectories on the street network is made through two steps: map matching and path reconstruction. The map matching is the process of locating the optimal placement of the GPS data on the street map (figure 2.3 left), while the path reconstruction is the process of

obtaining the optimal guess on the roads driven between two following placements (figure 2.3 right).

It has been developed[12] a map matching and a path reconstruction algorithm that work simultaneously, car by car, in order to perform the optimal map matching of the raw GPS data. This procedure takes place considering a particular car, that will have in our dataset a number N of recorded positions x_i and velocities v_i, where $i = 1, \ldots, N$. To every street arc j close to the measured position $x_i = (lat_i, lon_i)$, is assigned a matching probability $w_{i,j}$ with $\sum_j w_{i,j} = 1$. Each $w_{i,j}$ is a function of: i) the minimum distance between the measured position x_i and the arc j; ii) the angle between the speed v_i and the arc. The less are those quantities, the better will be the matching and thus the higher will be the matching probability. The weights may be summed $W_j = \sum_{i=1}^{N} w_{i,j}$ in order to describe a field W_j over the street network representing the number of times that the car has been located in a given street arc. The values of W_j, modulated by a sigmoid function, have been used to define a "discount" (up to the 20%) to the estimated (free flow) travel time cost. These discounted values T_j^f are then used to identify the shortest path for each possible map match couple $[(i, a), (i + 1, b)]$, with $w_{i,a} > 0$ and $w_{i+1,b} > 0$, of subsequent data records belonging to the same trajectory. Then, for each trajectory, the best paths are computed point by point beginning from the engine starting point matches and targeting the possible following target street arc. For all possible targets n, only one path, the shortest in terms of T_j^f, is then carried toward the next step. Paths that fail a consistency check with the distance covered (maximum tolerated error 10%) are excluded. If all paths from a to b are excluded, the trajectory is cut in a as if it would be a final target and a new reconstruction will start from point b. When we reach a final target, i.e. the engine stopping point or a cutting point, we may have different alternatives for the global path match of the whole trajectory. Coherently with the previous steps, we chose as global best path match the global path with lowest travel-time cost.

Chapter 3

Use of Time

When dealing with the modeling of human mobility, it appears natural to give time a central role. Time is spent traveling and time is spent in the destinations of our journeys, and it is people's common desire to minimize the time spent in traffic. The time available for our daily activities is limited by physiological needs like eating or going to sleep every night. Besides, working rhythms and personal habits shape each individual's timetable differently. On top of all that, even the more precise daily schedule may always be altered by some unexpected event.

It has been observed, from the analysis of the movements of bank notes[13], mobile phones[14] or private vehicles[15], that the time devoted to different locations follows a fat tailed distribution. This shows the fact that people spend most of their time in just a few locations while others are visited more shortly. The scaling properties of this distribution represents both one strong assumption for mobility models and one peculiar difference among data with different origins[17][18][16]. For these reasons, before starting with the actual characterization of human mobility, we start here determining the statistical features of time use when dealing with private vehicle mobility. Furthermore, taking advantage from datasets carrying information on time used also on the Internet and with mobile phones, we want to provide support to a more general approach in describing the perception of time.

3.1 Benford's Law of Activities' Duration

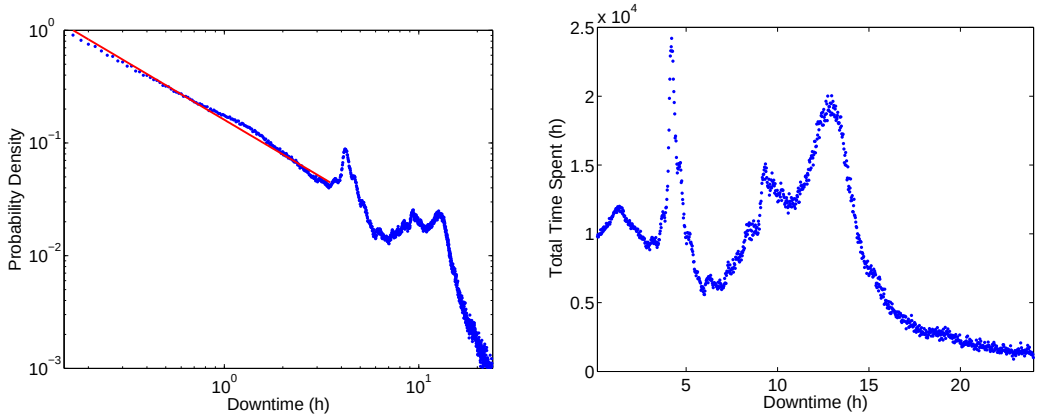


FIGURE 3.1: **(Left)** Statistical distribution of the activity times computed using GPS data of the Emilia-Romagna region (blue dots). The red straight line suggests the existence of a Benford's law $p(t) \propto 1/t$. **(Right)** Total activity time distribution (cfr. eq. (3.2)). The different peaks can be associated to the main individual activities: part-time job, full time job and the night rest.

GPS data do not give direct information on individual activities, but we may assume that each time a driver leaves the engine off more than 5 minutes, this can be associated to the execution of one activity. So, we can study the car engine's downtimes distribution, that we identify as activity durations, to understand how individuals use their time.

The result for the Emilia-Romagna dataset is plotted on the left of figure 3.1, where we point out the existence of a power law that accurately describe the distribution for $\tau \leq 3$ h ($\simeq 95\%$ of the data): a numerical interpolation of the experimental data gives $p(t) \propto 1/\tau^{\alpha}$ with $\alpha = 1.02 \pm 0.02$, and it is therefore statistically consistent with:

$$p(\tau) \propto 1/\tau \tag{3.1}$$

We recognize in this scaling law an alternative formulation of the Benford's law. In fact, the original Benford's law states that the probability of finding d as the first digit of given number N is $p(d) = \log[(d+1)/d]$. Pietronero et al.[19] have shown that the Benford's law emerges if the considered numbers are distributed following $p(N) \propto 1/N$, and that this distribution can be obtained in the limit $t \to \infty$ for a multiplicative process $N(t+1) = \xi N(t)$, i.e. multiplicative fluctuations as it happens in stock markets. Like a diffusion process gives a Gaussian, that for $\sigma \to \infty$ becomes an uniform distribution, the multiplicative diffusion gives a

lognormal, that for $\sigma \to \infty$ converges to $1/N$. For this reason we call equation 3.1 the Benford's Law of Activities' Duration.

It is important to remark that the empirical Benford's law for the time spent in the visited locations suggested by GPS data is not consistent with the analogous distributions computed from the mobile phone data[17] $p(\tau) \propto \tau^{-\beta}$ with $\beta = 1.8$; this can be the consequence of the finer time resolution of the GPS data, that allows to properly consider short time activities. GPS data suggest that distribution in fig. 3.1 is robust and does not depend on the spatial scale considered, as we have the same distribution considering different cities.

Differences among cities may instead be extracted considering the distribution $\pi(\tau)$ of the average time spent for activities with a time cost τ:

$$\pi(\tau) = \tau p(\tau) \tag{3.2}$$

This quantity is constant where the profile of the downtime probability density follows exactly the statistics of equation 3.1. Figure 3.1 (right) shows instead a peak structure, a signal modulating the underlying Benford's law statistics. The most prominent peaks are related to the main human activities: the part time job (rest time $\tau \simeq 4$ h), the full time job (rest time $\tau \simeq 8$ h) and the night rest. Another significant peak is also around $\tau \simeq 1.5$ h, while others smaller may be identified around multiples of one hour. The relative heights, the width and the position of those peaks in different analyzed areas may be compared in order to investigate over the peculiarities of distinct areas, because this profile is a direct consequence of the daily habits of each city's inhabitants.

3.1.1 Progressive time usage model

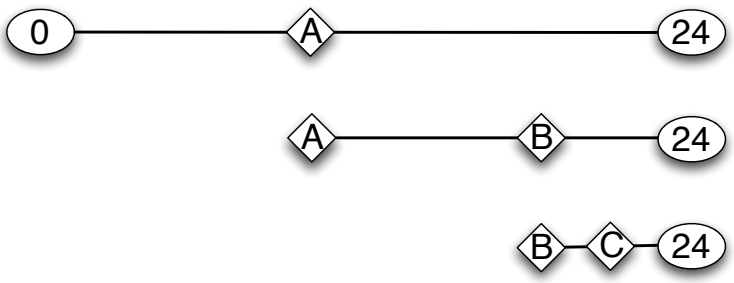


FIGURE 3.2: Graphical representation describing the first three random choices of activity lengths, compatibly with previous made choices, given an initial 24 time budget.

In order to give a microscopical interpretation of the empirical downtime distribution, we propose here a simple model showing how the observed Benford's law may be a consequence of a progressive scheduling of a limited daily time budget. This approach is antithetical to spontaneous behavior[20]: as individuals make plans, the programmed future activities influence the duration of the preceding ones.

For sake of simplicity we identify a fixed temporal constraint in the circadian rhythm. Therefore, the total daily time budget, which can be distributed to the many activities performed in a day, is assumed to be of 24 hours. This choice may be recognized as a micro-canonical approach, where the energy (in this case the total daily time budget) is almost exactly determined. Moreover, ignoring the real habits of the individual, we assume that they cannot precisely determine a priori each activity downtime, because this is varied depending on unpredictable circumstances.

Under these conditions, each individual progressively consumes the time budget with a succession of random choices (see figure 3.2). First choices will take the biggest portions of time, mimicking the role of the main activities such as night rest and work. The following choices can allocate less and less time, as most of it has already been assigned, and thus they represent the way the free time between a fixed activity and another can be spent.

If one computes the interval distribution that is obtained by the stochastic process of choosing successively k points in a given segment as in fig. 3.2, one gets analytically the Benford's distribution (for the analytical proof, see appendix A); this result has been also verified with Monte Carlo simulations.

3.2 Relationship between Activity Time and Degree

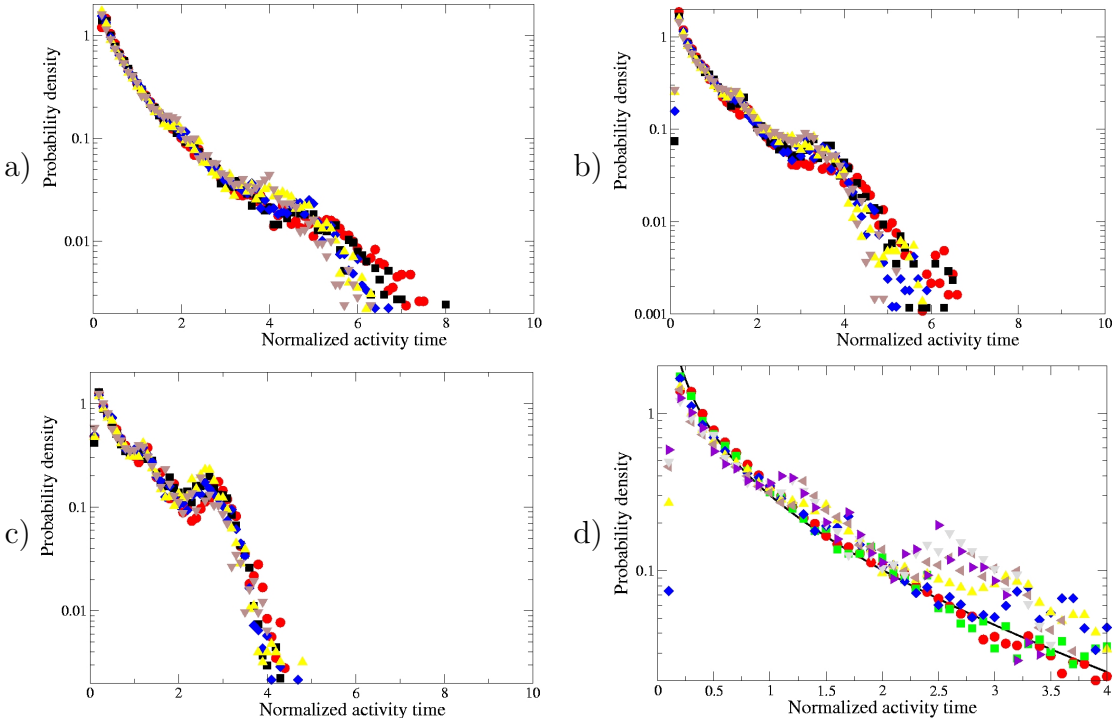


FIGURE 3.3: Empirical distributions in the Florence dataset for the conditional probabilities $p(\tau \mid k)$ for activities performed $k = 3, .., 20$ times in a month as a function of the normalized activity downtime $\tau/\langle\tau\rangle_k$. The different symbols refer to the different activity degrees. **a)** k = 3 (circles), k = 4 (squares), k = 5 (rhombus), k = 6 (up triangles) and k = 7 (down triangles); **b)** k = 8 (circles), k = 9 (squares), k = 10 (rhombus), k = 11 (up triangles) and k = 12 (down triangles); **c)** k = 16 (circles), k = 17 (squares), k = 18 (rhombus), k = 19 (up triangles) and k = 20 (down triangles); **d)** Combined representation of various of k ranging from 3 to 20: the continuous line refers to an interpolation with the function (3.4).

The Benford's law for activities' downtime outlines the stochastic features of the system, but it does not explain how such features can be related to the individual daily agendas, which are certainly the result of a cognitive behavior. In order to study this question, we perform a statistical analysis of the downtimes related to the different individual activities that present the same monthly degree k (i.e. the number of times that a citizen repeats a certain activity during a month)[21].

Let τ the activity downtime, we introduce the joint probability $p(\tau, k)$ to denote the probability of finding a k-degree activity associated to a downtime τ. Then,

by definition we have to recover the Benford's law by summing over k:

$$\sum_k p(\tau, k) \propto \frac{1}{\tau} \tag{3.3}$$

We have also the equality $p(\tau, k) = p(\tau \mid k)kp(k)$ where $p(\tau \mid k)$ is the conditional probability for a downtime t considering only the k-degree activities, and $p(k)$ is the probability to detect a k degree activity; the factor k takes into account the multiplicity of the k degree activities. The study of the conditional probability $p(\tau \mid k)$ can shed some light to understand the mobility habits related to the use of private vehicles and to face the question of the relevance of repeated activities both in the mobility and in the use of time.

In figure 3.3 we plot the empirical probability densities for different degrees (from $k = 3$ to $k = 20$), to investigate the existence of the universal distribution $f(u)$. There is a decreasing of the data number as k increases, but all the distributions are computed with a sample of the same order (from 4×10^4 to 10^4). The figures enlighten three different features. There is a collapse of all the curves on a unique distribution: this is clear in the figure 3.3-**a** (the tail spread is consistent with statistical fluctuations) and in the first part of all the plotted distributions that contains the great majority of the data. All the distributions show a big contribution from the short times activities and a fast decaying tail for large $(\tau/\langle\tau\rangle_k) > 2$. There is a smooth rise of a "signal" as k increases denoted by the appearance of two peaks at $\tau/\langle\tau\rangle_k \simeq 1$ and $\tau/\langle\tau\rangle_k \simeq 3$: this is clear in 3.3-**c**. Therefore the empirical observation gives a strong indication for the existence of an universal distribution $f(u)$ for the normalized activity downtime, even if when we consider high degree activities ($k \geq 10$) some new features appear but with a small statistical weight. A possible interpolation of the distribution $f(u)$ is given by:

$$f(u) \propto \frac{1}{u} e^{-\alpha u} \tag{3.4}$$

where the coefficient α has a value $\simeq .4$. The distribution (3.4) is singular at the origin so that the interpolation is certainly approximated at $u \to 0$ (see fig. 3.3-**d**).

Remarkably, these experimental observation suggests the existence of an universal probability distribution $f(u)$ for the normalized downtime $\tau/\langle\tau\rangle_k$:

$$p(\tau \mid k) = \frac{f(\tau/\langle\tau\rangle_k)}{\langle\tau\rangle_k} \tag{3.5}$$

where $\langle\tau\rangle_k$ is the average downtime for the k-degree activities. We read this universal function as the signature of the fact that individuals organize their time, when performing a private car mobility, in a common way independently from the specific activity, i.e. the relative downtime fluctuations are the result of a stochastic universal mechanism. Moreover, there should exist a common feature among the individuals, concerning how they manage the downtime related to the k-degree activities, since only the average value $\langle\tau\rangle_k$ characterizes the k dependence of the conditional probability $p(\tau \mid k)$. This universal character could be explained thinking that the $\langle\tau\rangle_k$ variable is a "measure" of the mobility actions, valid for every individual. More precisely, $\langle\tau\rangle_k$ can be considered the temporal norm for all the mobility related activities. From the empirical data in the Florence dataset we detect $\simeq 3 \times 10^5$ activity downtimes and we have computed the dependence of the average value $\langle\tau\rangle_k$ using the degrees $k = 3, .., 20$.

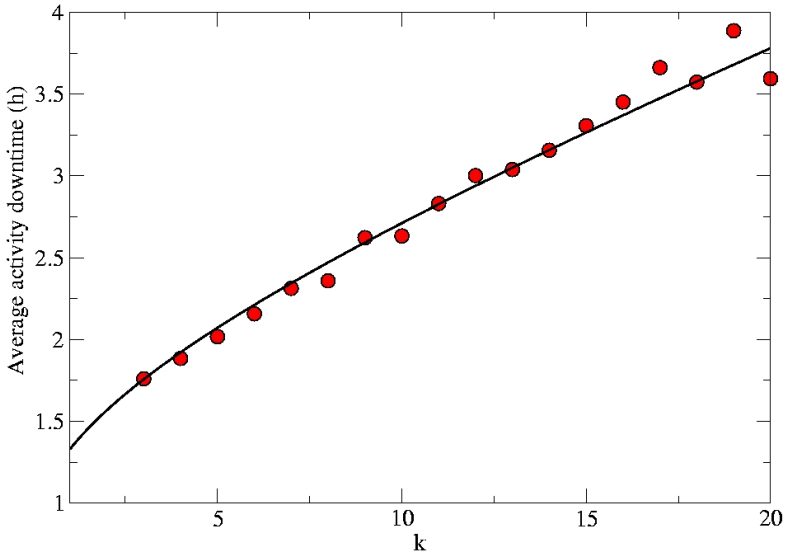


FIGURE 3.4: Dependence of the average downtime $\langle\tau\rangle_k$ from the activities' degree k in Florence. The continuous line refer to a possible interpolation with the exponential function (3.6).

From the result, shown in fig. 3.4, appears evident that we have an almost linearly increasing behavior of $\langle\tau\rangle_k$ as the degree k increases. This means the existence of a relation between the activity degree and the activity "use value" (individual satisfaction, profit, etc...) introducing an individual tendency to repeat and to spend time in the activities with a relevant added value[22].

A possible local interpolation of the empirical data is obtained by using the function (continuous line in fig. 3.4):

$$\langle \tau \rangle_k \propto \exp(\gamma k^a) \tag{3.6}$$

where $a \simeq .3$ and $\gamma \simeq .7$.

3.3 Weber-Fechner's Law of Perceived Durations

Benford-like distributions have been observed in other works on time related behavior. Some examples are: the time between consecutive emails[23], phone calls durations[14], visits of a web portal or library loans by a single user[24]. A queuing model has been proposed[23] to explain the nature of inter-event time distributions, but which was the correct interpolation's curve for the distribution's heavy tail has been object of debate[26][25]: depending upon the methodology either a lognormal[25]:

$$p(\tau) = \frac{1}{\tau\sqrt{2\pi\sigma^2}}\, e^{-\frac{(\ln\tau-\mu)^2}{2\sigma^2}} \tag{3.7}$$

or a power law with exponent -1 and an exponential cutoff[24]:

$$p(\tau) \propto \frac{1}{\tau} e^{-\frac{\tau}{\tau_0}} \tag{3.8}$$

were endorsed.

Therefore, we have focused our attention on the tail of our activity time distribution, switching from a linear binning as in fig. 3.1, where for times greater than 24 hours the distribution would have resulted too noisy, to a logarithmic binning that produces figure 3.5. There, for values of the activity time ranging from 2 minutes to 9 hours, we recognize the Benford's Law, where for values over 9 hour the interpolation has been made with a lognormal function. The distribution's tail clearly overcomes the limited budget of 24 hours assumed for the progressive usage model, and is therefore related to stop durations out of the range that cannot be directly described by our model without introducing fluctuation in the total activity time budget.

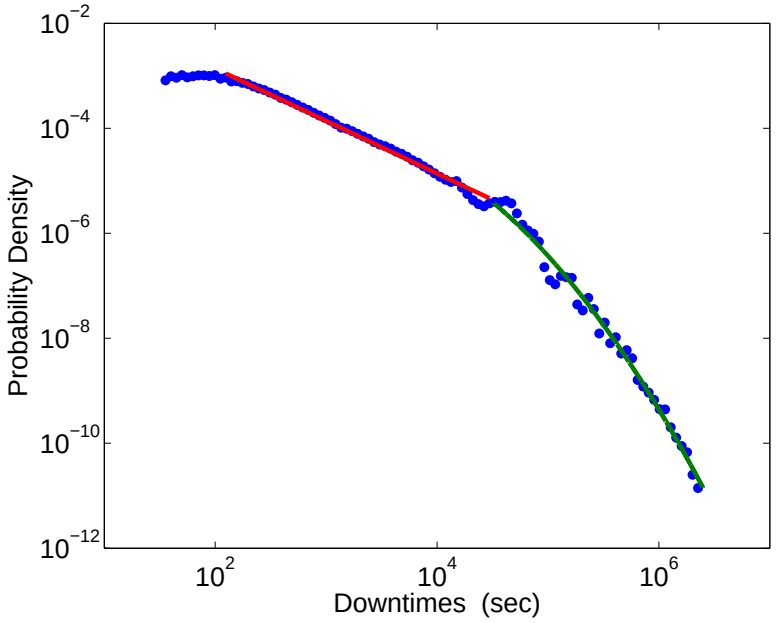


FIGURE 3.5: Statistical distribution of activity times computed using GPS data of the Italy dataset. The red line represents an interpolation with a power law $p(\tau) \propto \tau^{-1.04\pm0.03}$ for $\tau \in$[2 minutes, 9 hours] and the green line an interpolation with a lognormal for τ greater than 9 hours.

Then, we tried to apply a similar approach to inter-event times relative to mobile phones, taken from the Reality Mining dataset[27]. In this case has been possible to fit (see fig. 3.6 left) both the short and long times tails with only one lognormal curve, which parameter μ has been fixed to the empirical average value of $\log \tau$ in the sample and only σ is a free parameter. At the same time, in the interval within 30 seconds and 12 hours the distribution follows closely the Benford's Law.

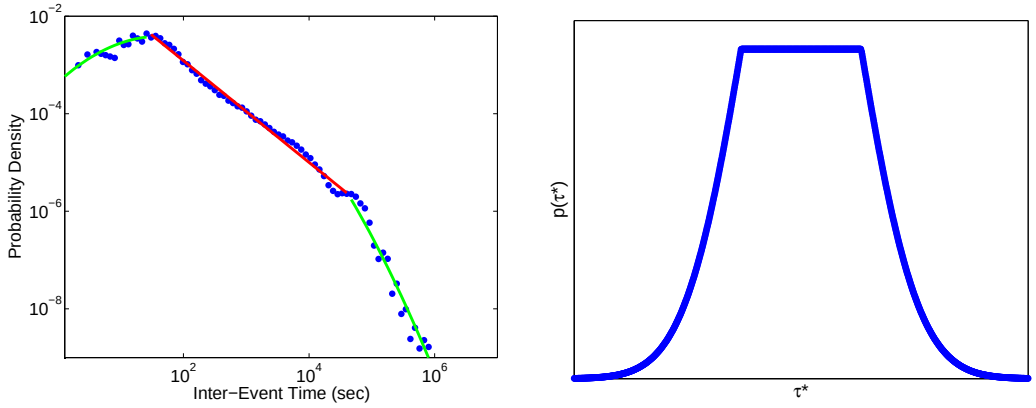


FIGURE 3.6: **(Left)**Mobile phone inter-event times distribution. Within the interval 30 seconds - 12 hours the fit is with a power law with exponent -1.00 ± 0.03, while outside this range with a lognormal distribution. **(Right)** An illustration of the proposed distribution of $\tau^* = \ln \tau$, with the coexistence of two maximum entropy distribution: uniform and normal.

Furthermore, we have obtained a new independent dataset that enable us to proceed with another investigation on activity times. We have been given the rights to export from the Google Analytics data regarding the image bookmarking website *imggot.com*[28]. Analytics data give every hour the number of visitors and the average duration of the visits. Considering only the hours when only a single visit occurred we have extracted from 3 years of recording ≈ 2000 duration times, that are distributed as in figure 3.7. So, we have a perfect second example of Benford's Law of duration times, that might be again interpreted as a consequence of a progressive consumption of a finite time budget (as shown in section 3.1.1).

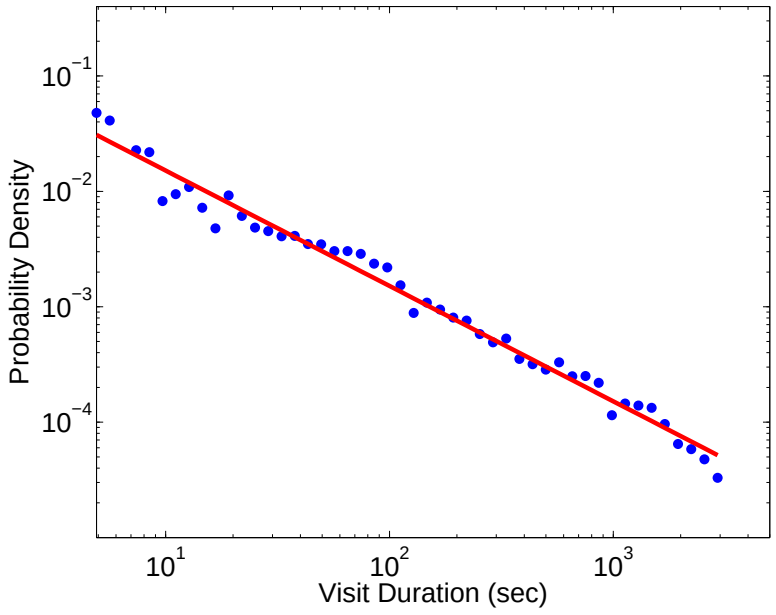


FIGURE 3.7: Distribution of the visit durations on *imggot.com*. The red line shows the best fit with the power law $p(\tau) \propto \tau^{-1.00\pm0.04}$.

The log-normal tails suggest that, alternatively to both the time consumption model and the queuing model for inter-event time, we can take into consideration a logarithmic time perception as the root of both the Benford's Law and the lognormal cutoff. In fact, with the change of variable:

$$\tau^* = \ln \tau \tag{3.9}$$

the Benford's Law becomes an uniform distribution (on an interval), while the lognormal distribution becomes the normal distribution. Both normal and uniform distributions are maximum entropy distributions: the first among the real-valued distributions with assigned mean and standard deviation, the second among the continuous distributions supported in an interval. Therefore, under those and the log-time perception assumptions, there is no real need for a model to explain these curves. The change in behavior from one to the other curve suggests that time intervals belonging to the range of validity of the Benford's Law are equivalent and different by those outside that range, where τ^* is limited by the need of having average and variance and thus follows a Gaussian distribution. An idealized distribution for τ^* is represented in figure 3.6 right.

Logarithmic time perception is one aspect of the psychophysical Weber-Fechner's law. Numerous human responses to physical stimulus of quantities such as brightness, loudness and weight appear to be naturally compressed in a logarithmic

encoding. This logarithmic scale has the advantage of permitting a compact representation that can cover several order of magnitudes with a constant relative error[29]. As an alternative to Weber-Fecher's law has been proposed the Stevens' law, stating that the relationship between stimulus and responses is better modeled by a power law. Nevertheless, as experimental results report that in general, power relations for duration have coefficients in the interval 1 ± 0.1[30][31][32], in this case Stevens' law would be almost equivalent to linear response. But, these studies were focused only short time perceptions. When dealing with times longer than 5 seconds the evaluation of a duration involves memory and may be influenced by expectation and attention[32][1].

Concerning longer periods, it has been observed that human temporal cognition in inter-temporal choice in the range one week-25 years follows the Weber-Fechner's law rather than Stevens power law[33]. The logarithmic perception of temporal duration might also explain the hyperbolic discounting of delayed rewards, that can be put in relationship with drug-addicted patients' behaviors[34] and consumers' decisions[35].

[1] Indeed, for longer time interval it would be more correct to use the term "time estimation"[32].

Chapter 4

Trip Lengths and Mobility Budgets

In this section, we begin the statistical characterization of human mobility focusing on the spatial characteristics of trips. The most important quantity for this analysis is the distance covered while traveling. Understanding the relation between trips lengths, urban structure and economical indicators, represents a necessary step in order to reduce energy and environmental problems and converge to sustainable cities[1].

In the following we propose a theoretical explanation for the statistical distribution of trips lengths, based on the assumption of the existence of a daily Mobility Budget that is linked to a Total Travel-time Budget. The specific relationship between a trip length and its duration will be investigated in one of the next chapters (section 7.1).

4.1 Curvilinear-Euclidean Lengths Relationship

Before moving forward to the analysis of the statistical properties of trip lengths, it is essential to define a method to evaluate this quantity. In fact, in our data are available three different measures of length (see fig. 4.1):

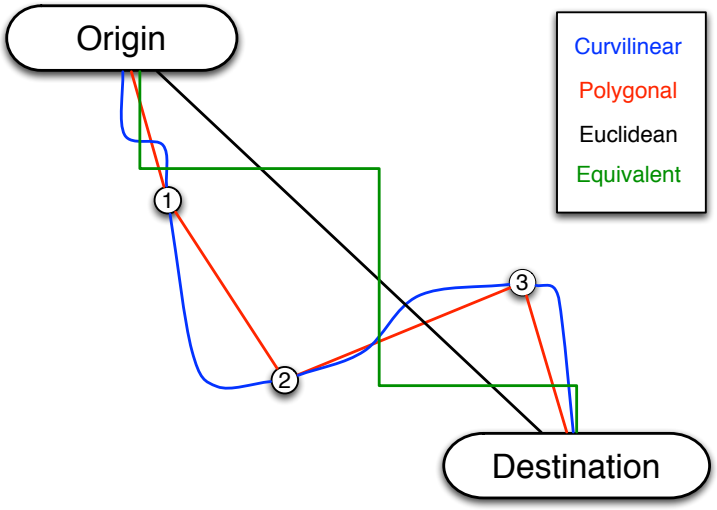


FIGURE 4.1: Different types of distance associated to a trip where starting datum, stopping datum and 3 travel data have been recorded.

Euclidean distance: distance "as the crow flies" between starting and stopping point;

Polygonal distance: sum of the euclidean distances between the subsequent GPS positions recorded for the same trip;

Curvilinear distance: sum of the "distances from the previous datum" fields for data in the same trip, where the GPS device has progressively added the distances between the intermediate GPS positions that are not recorded in memory.

Each of these measures has advantages and disadvantages. Curvilinear distance is in principle the most precise, but if the most influenced by errors due to week or absent GPS signal. Instead, polygonal and euclidean distances are systematically underestimating the real values of distance covered, but are less subject to errors.

These differences are evident when we see the probability densities of the three types of distance in the Florence dataset (fig. 4.2). Polygonal and euclidean distances are in good correspondence after that the euclidean distances are multiplied by a factor $\sqrt{2}$ (quantity indicated in figure as "equivalent length). Curvilinear lengths are greater than polygonal ones, and this difference is especially remarkable for short trips. On the other hand, the peaks in the curvilinear distances distribution, for multiple values of 2 Km, are probably consequences of signal losses. What we suppose it is happening, as 2 Km is intra-records distance imposed by

the GPS device design, is that if between this recording and the trip's end the device lose contact with the satellite, curvilinear distance cannot be updated and the last part of the trip is missed, while the stopping position used in the other measures is subject to a correction process reducing this type of errors. This correction grants also an average a better quality to at the starting and stopping point with respect of data taken during movement. As we have observed that, in the Florence dataset, a fraction of curvilinear and polygonal distances pathologically wrong, such as negative values of curvilinear distance, we have chosen euclidean distance as reference measure as it is the most trustworthy, being calculated only using starting and stopping positions

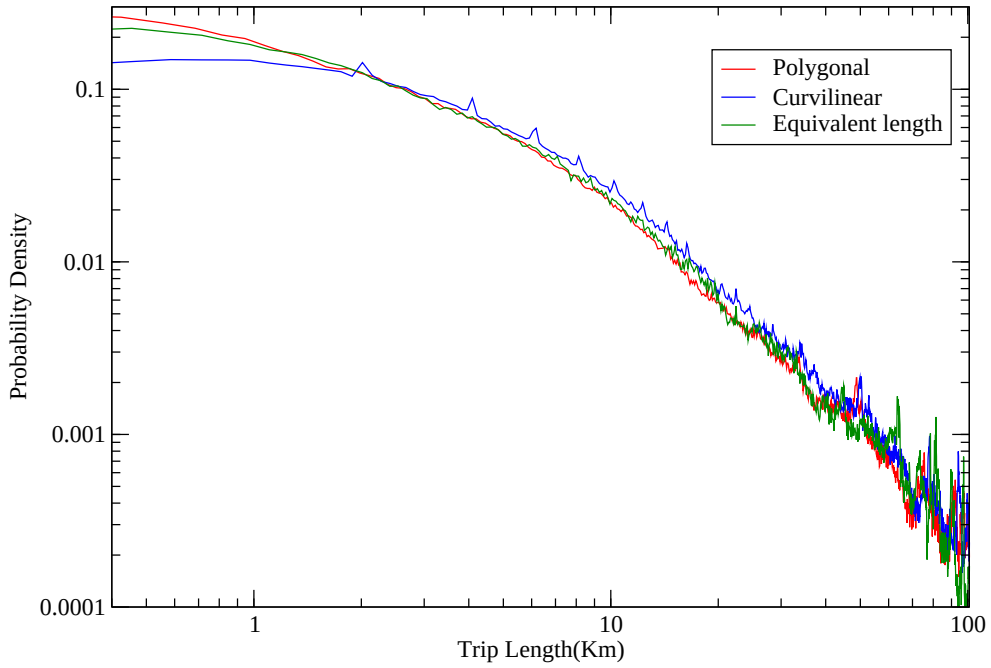


FIGURE 4.2: Probability densities of polygonal and curvilinear distance, confronted with the equivalent distance l proportional to the euclidean.

Therefore, taking advantage of the euclidean distance d_e we use an equivalent distance:

$$l = \sqrt{2} \cdot d_e \tag{4.1}$$

that is in a good statistically correspondence with both polygonal and curvilinear length, which are by nature a good approximation for the effective length of the

vehicle's trajectory. We can imagine this new quantity as the measure of a fictional, zig-zagging trajectory constituted by a sequence of orthogonal sections (see fig. 4.1)[1]. This equivalent distance is able to statistically substitute the constraint represented by the street network that does not permit to the drivers to direct towards the destination.

From now on and where not differently specified, trip lengths are intended as measure of the euclidean distance d_e, and if a comparison with actual distance covered is needed, that can be made using the equivalent distance l defined through the relationship (4.1).

4.2 Trips Length's Distribution

We consider here the trip lengths distribution (figure 4.3), computed using the Emilia-Romagna GPS data. In this particular case, trip have been considered completed when the rest time is longer than 5 minutes, otherwise we sum the lengths between two successive stops.

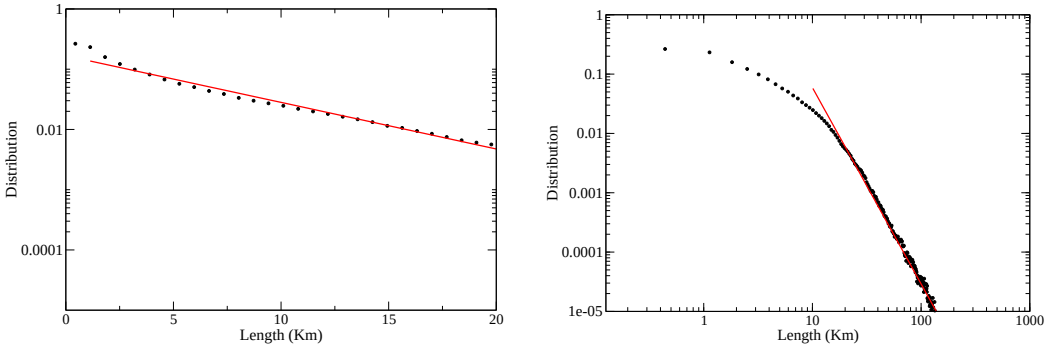


FIGURE 4.3: Statistical distribution of the trip lengths for the Emilia-Romagna: we use log-lin scale in the left plot and log-log scale in the right plot. The log-lin scale suggests a possible exponential behavior for the short trips, which represent the 95% of the data (red line). The log-log scale points out a possible interpolation of the distribution tail by a power law: $p(l) \propto l^{-3.3}$ (red straight line).

We remark on three main features:

- the very short trips ($l \leq 2$ km) have a great statistical relevance;

[1] We may notice that this virtual path is similar to a measure of length in a "taxicab geometry" metric. Although, we have that our equivalent distance is $\geq L_1$ distance.

- there exists a characteristic trip length $\simeq 6.2$ km;
- the long trip distribution recalls a fat tail (power law) distribution.

The trip lengths distribution reflects the way everybody realizes his mobility demand in connection with the spatial activity distribution[36]. In the following, we propose a theoretical explanation for this distribution.

4.2.1 Individual Mobility Budgets

As a consequence of the circadian rhythms, it is natural to consider the daily mobility as limited for both for physiological and economical reasons (any trip has a cost in time, energy and money). Thus, we can define a quantity λ for each individual and each day of mobility, defined by the sum of the trip lengths of performed within an interval of 24 hours. This quantity describes the total daily length covered and we will call it "Daily Mobility Length" or just "Daily Mobility"[2].

The daily mobility distribution computed from the GPS data is plotted in fig. 4.4 together with an exponential interpolation.

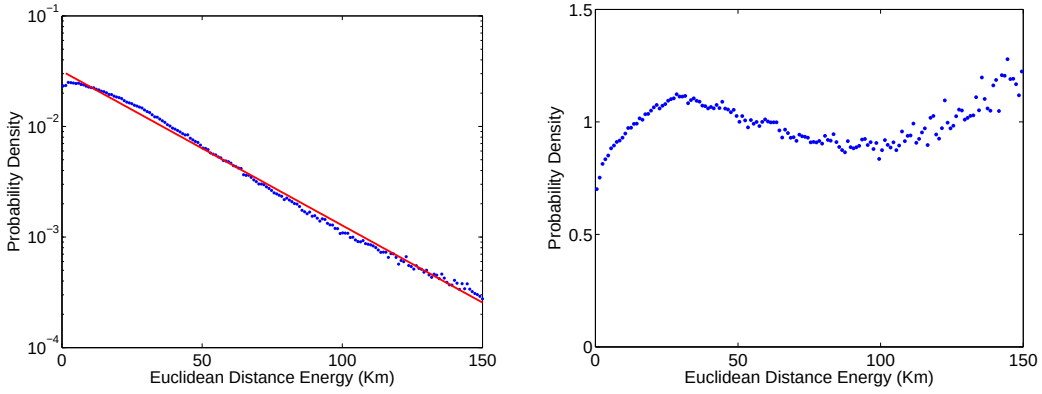


FIGURE 4.4: **(Left)** Daily mobility distribution from the GPS data selecting people moving inside the Emilia-Romagna region. The straight line refers to an exponential fit of the distribution with a characteristic length $\beta^{-1} = 30.4 \pm 0.4$ km. **(Right)** $m(\lambda)$ distribution (cfr. definition (4.6)) computed using (4.5).

[2]The term "Energy" has also been used by us in [15] and in some figures of this dissertation, both because an underlying energy consumption concept has been been suggested in [37] and because the probability density of this quantity follows the Maxwell-Boltzmann energy distribution.

We suggest a theoretical explanation for this behavior by dividing the territory into a number of different locations $x \in X$, with homogeneous geographical features. Assuming a given activity distribution in the territory, we associate to each location a daily mobility length λ_x defined by the average distance that an individual has to cover each day to satisfy his mobility demand (in other words λ_x measures the accessibility of the x-location to the existing activities). Let p_x be a priori probability that an individual chooses to live in the x-location without taking into account any mobility cost[3]. Assuming that individuals act as independent particles, the probability associated to a distribution $\{n_x\}$, where n_x is the number of individuals in the location x, is given by a multinomial distribution:

$$w(\{n_x\}) = \prod_x \left(\frac{p_x^{n_x}}{n_x!} \right) \tag{4.2}$$

Applying a maximal entropy principle with the constraints that the total number of individuals and the total mobility are finite:

$$\sum_x n_x = N \qquad \sum_x \lambda_x n_x = \Lambda$$

one can determine the most probable distribution. Maximizing the Gibbs entropy[38]:

$$S = -\sum_x w(n_x) \ln w(n_x) \tag{4.3}$$

we get the Maxwell-Boltzmann distribution

$$\rho(x) = A \exp(-\beta \lambda_x) p_x \tag{4.4}$$

where A is a normalizing constant and β depends on the average mobility $\beta^{-1} = \Lambda/N$. Adding over all the locations with the same value $\lambda_x = \lambda$, we finally get the distribution:

$$\rho(\lambda) = A\, m(\lambda) \exp(-\beta \lambda) \tag{4.5}$$

where

$$m(\lambda) = \sum_{\lambda_x = \lambda} p_x \tag{4.6}$$

The measure $m(\lambda)$ gives the statistical weight of individuals that would perform a daily mobility λ, if their distribution in the territory does not depend on mobility costs related to trips lengths. As shown in fig. 4.4, the daily mobility distribution is

[3] In a homogeneous territory p_x would be constant, otherwise p_x may depend on the geographical features.

quite well interpolated by an exponential distribution in the interval 10 km $< \lambda <$ 150 km. The distribution $m(\lambda)$ estimated according to the formula (4.5) (figure 4.4 right), has a limited variation within this interval with a local maximum at $\lambda \simeq 30$ km, that reflects the macroscopic spatial distribution of activities in the Emilia-Romagna territory. Therefore a possible explanation for the $m(\lambda)$ behavior is the following: considering that activities are mainly located in the cities, the initial increase of $m(\lambda)$ is due to the population living in the attraction basin of the cities and the maximum at $\simeq 30$ km gives an estimate of the average distance among the main cities.

The statistical distribution (4.4) leaves open the question if the exponential decay is related to the extension of the considered region. Then, we have compared the daily mobility related to areas of different size R centered on Bologna (the regional capital), from the Bologna province ($R \leq 30$ km), to the area enclosing the nearby cities ($R \leq 50$ m) and then to the whole region. In each area we have only considered individuals whose mobility is performed internally to the area itself, but that have not been previously (for smaller radius) considered. We recall that our analysis refers to the use of private vehicles and we expect cars to be utilized to satisfy the same mobility demand in all the cases; this is false inside urban areas ($R < 5$ km) where one has a good availability of public means and more restriction in the use of private cars. The resulting distributions are reported in fig. 4.5 where the exponential decaying can be clearly detected at different scales, and for large daily mobility we see a different behavior close to the main city.

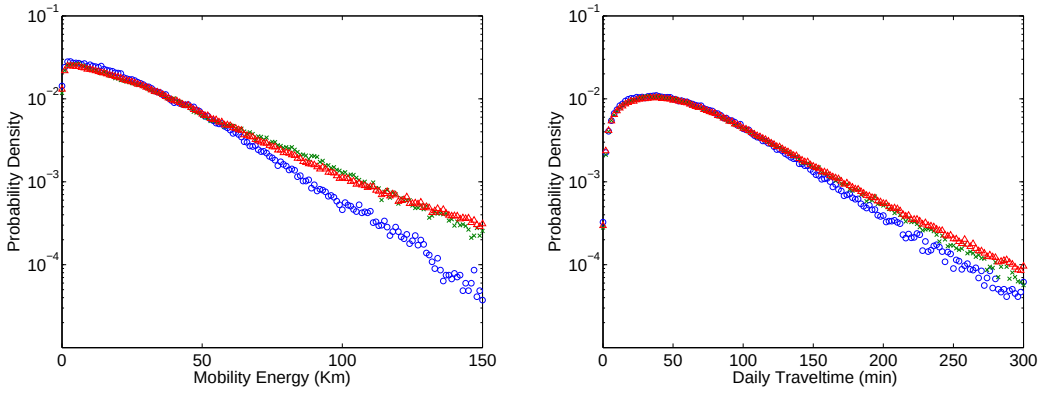


FIGURE 4.5: **(Left)** Daily mobility distributions computed considering individuals performing their mobility inside regions of different size around the Bologna center: the circles refer to the Bologna province $R \simeq 30$ km, the crosses refer to region that includes the nearby cities $R \simeq 50$ km and the triangles give the distribution for the whole region. **(Right)** Daily travel-time distribution corresponding to the daily mobility distributions plotted in the left picture: the symbols have the same references as in left picture.

The results suggest that the entropic principle is robust in describing the average mobility demand, but the characteristic spatial scale decreases approaching an urban area.

Furthermore, in transportation planning and modeling is frequently taken as key concept the Travel Time Budget. This quantity, measuring the time that every day an individual accepts to invest in his mobility, has been assumed as an universal constant of $\approx 1.2 \pm 0.1$ hours per traveler per day[39].

Considering the travel-time budget distributions on the previous regions of different size (see figure 4.5 (right)), they tend to collapse into a single curve. Interestingly, the average mobility time is estimated over all the Emilia-Romagna region is 70 minutes (1.17 hours) from our GPS data. This experimental evidence seems to support the definition of a universal cost for mobility (once the transportation mean is given). We will show in the following section 4.3 that, when we focus on the urban scale, the Italy dataset suggests a dependence of the travel time budget from the city where the travelers live.

Comparing the figures 4.5 left and right, we remark that the space-time relationship cannot be reduced to a simple proportionality. The reason is twofold: from one hand there is an intrinsic heterogeneity in the human mobility due to different drivers behaviors, on the other hand the small scale structure of the road network influences the vehicle dynamics. A specific study on the relationship between trip length and duration has been made in section 7.1.

4.2.2 Random Partitioning of the Daily Mobility Length

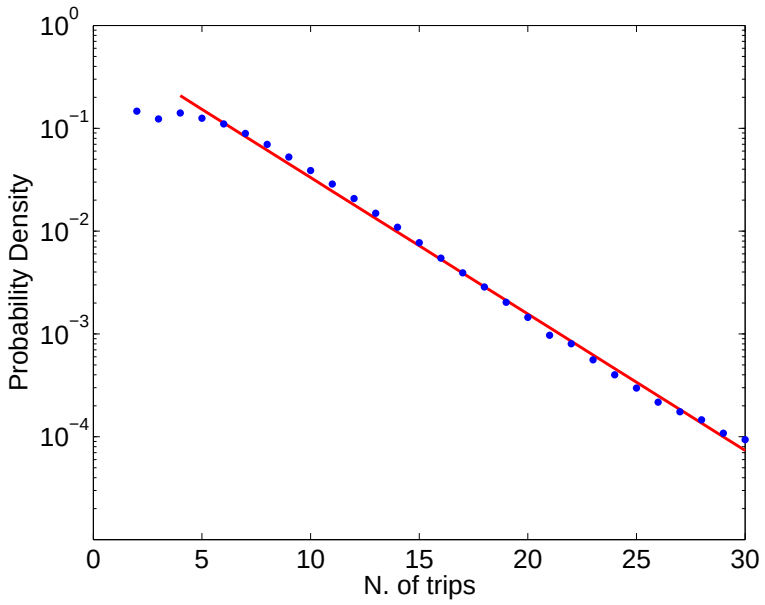


FIGURE 4.6: Distribution of the daily activities number for the sampled individuals in Emilia-Romagna; the continuous line is an exponential interpolation $p(n_t) \propto \exp(-n_t/a)$ with $a = 3.27 \pm .08$.

In order to relate the trip lengths distribution (fig. 4.3) with the daily mobility (fig. 4.5), we have to consider how many trips each individual makes in a day. In figure 4.6, we plot the probability distribution of the trips number together with an exponential interpolation. For $n_t \leq 5$ we have about half of the sample population that presents an almost uniform activities number distribution. Probably, this sample represents people who are performing a systematic mobility involving a few essential locations such home, workplace and grocery stores. Besides, where $n_t > 5$, the exponential decay suggests a statistical equilibrium without any particular structure in the individual mobility. Indeed, the emergence of an exponential distribution is consistent with the assumption that, in average, individuals behave as independent random particles that define their daily agenda in a random way.

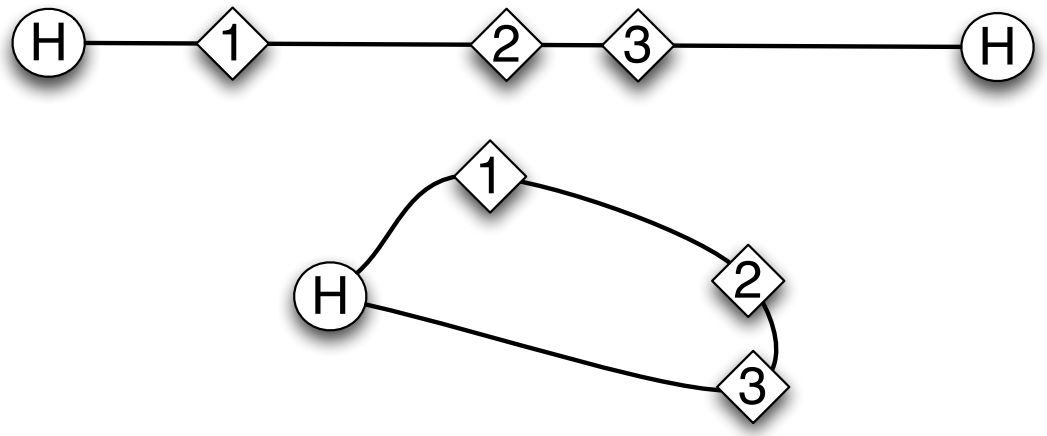


FIGURE 4.7: Graphical representation of the distribution of stops in a segment representing the mobility length λ of a day where in agenda there are $n_t = 4$ trips ($n = 3$ stops) and both nights are spent at home.

To interpret the statistical part of the trip lengths distribution (cfr. fig. 4.3), we consider an ensemble of particles characterized by a total mobility λ and, for each particle, we randomly distribute at most $n = n_t - 1$ destinations within the interval $[0, \lambda]$, as pictured in figure 4.7. The obtained distances among the destinations are the trips performed by individual-particles. Given λ and n, the trip lengths distribution can be computed analytically according to:

$$p_{n,\lambda}(l) \propto \sum_{k=1}^{n} e^{-k/a}(k+1)k(1-l/\lambda)^{k-1} \tag{4.7}$$

(see Appendix B or [15] for the derivation of the last formula) where $l \in [0, L]$ and $a = 3.27$ is determined from the empirical activities number (see fig. 4.6). Using the exponential distribution (4.5), where we neglect the changes due to $m(\lambda)$, and integrating over λ, we get an analytic formula for the trip lengths distribution:

$$p_n(l) \propto \int_{\lambda_m}^{\lambda_M} p_{n,\lambda} \exp(-\beta\lambda) d\lambda \tag{4.8}$$

In the fig. 4.8 we compare the empirical trip lengths distribution with our analytical result (4.8).

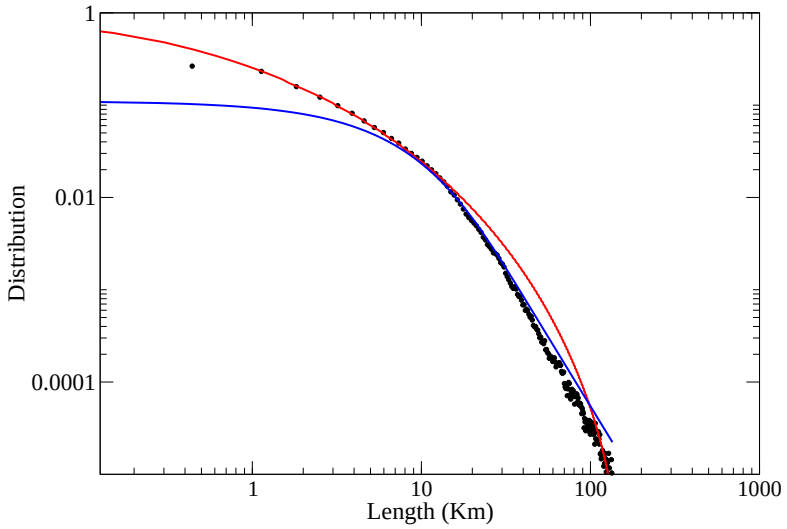


FIGURE 4.8: Statistical distribution of the trip lengths measured using GPS data (black dots): the red curve refers to the distribution (4.8), whereas the blue curve is computed using eq. (4.9) with $\lambda = 180$ km and $n = 30$.

Formula (4.8) closely interpolates the experimental data for short trip lengths $l \leq 15$ km[4]. This allows to reproduce the mobility of $\simeq 87\%$ of the observed users, that correspond to $\simeq 52\%$ of the total space traveled. But, we have a discrepancy in the tail of the empirical distribution. A possible explanation is obtained if one does not introduce the exponential decaying in the number of trips (cfr. fig. 4.6), so that the distribution (4.7) reads:

$$\begin{aligned} p_{n,\lambda}(l) &= \frac{2\lambda^2}{n(n+3)} \sum_{k=1}^{n} (k+1)k(1-l/\lambda)^{k-1} = \\ &= \frac{2\lambda^3}{n(n+3)} \frac{d^2}{dl^2}(1-l/\lambda)^2 \left(\frac{1-(1-l/\lambda)^n}{l} \right) \end{aligned} \tag{4.9}$$

Since $1 - l/\lambda$ is small for long trips ($l \simeq \lambda$) for $n \gg 1$ we can approximate:

$$p_{n,\lambda}(l) \simeq \frac{2\lambda^3}{n(n+3)} \frac{d^2}{dl^2}(1-l/\lambda)^2 \frac{1}{l} \propto l^{-3} + O(\lambda^{-2}) \tag{4.10}$$

The power law tail (4.10) seems to be in agreement with the empirical observations (see fig. 4.8 for a comparison of (4.9) with the experimental data). This result suggests that we have users with a number of trips higher than the statistical expectation and with a large mobility. This can be due to a correlation between

[4]The discrepancy at very small trip $l < 1$ km is expected since using the exponential distribution $e^{-k/a}$ for the activities, small trips are overestimated.

λ and n_t in a subset of vehicle, probably driven for working reasons. We remark that the power law $p(l) \propto l^{-3}$ is different from the power-laws suggested by the dollar bill displacement distribution $p(l) \propto l^{-1.59}$[13] or by the mobile phone data $p(l) \propto l^{-1.55}$[17]. But both consider a much larger spatial scale and do not refer to a particular transportation mean.

4.3 Travel Time Budget

Transportation research on travel times is dominated by the assumption of the existence of a constant travel-time budget[40]. This concept is based on the behavioral hypothesis that people have a certain amount of time that they are willing to spend on travel, and they therefore tend to minimize deviations from that budget in either direction while maximizing the utility coming from the activities carried out at the destinations[41]. Drivers will therefore trade travel-time savings for more trips to perform[40][5]. In addition, also the share of monetary expenditure that individuals allocate to transportation (money budget) has been indicated as one decisive factor for aggregate travel behavior. Both travel time and money budgets have been suggested to be characterized relatively stable distribution that appear to be universal if aggregated for different cities and times[39].

Although, a slight tendency to travel time budget increase with the size of the area was already been observed[40], and budget in the same nation has been recently been observed to be growing over time[42]. A meta-analysis in different aggregate and disaggregate studies of travel time expenditures has suggested that they are also strongly related to the characteristics of the individual, of the destination activities and of residential areas[41]. Moreover, a constant travel-time budget is only consistent with empirical data when considering only a single mode of transport. Conversely, energy consumption rates specific for each form of transport have to be taken into account to define an universal travel-energy budget[37].

In the last part of this chapter, we show an analysis of the travel-time budget distribution, aggregated over different cities in our Italy dataset, where we find noteworthy dependencies on either census or traffic related quantities. We integrate this subject in the chapter on trip lengths, as travel time budget is evidently

[5] Is attributed to Zahavi the observation that, rather than assuming the individual to be asking, "What is the least amount of travel I can do in order to accomplish a given set of activities?", the individual instead should be viewed as asking, "What is the most attractive set of activities/destinations I can achieve, given a certain travel time budget?".

related with total mobility, that we have seen is a key quantity to explain the trip lengths distribution.

4.3.1 City Dependency

From our Italy dataset, we have systematically extracted the average values of daily travel times for 1233 cities where we have at least 100 drivers, a sample statistically dominated by medium sized and small municipalities. In this particular analysis we have defined one driver as belonging to a city if the most part of its parking time was spent in the municipality area. Then, all the mobility performed in a day (in and out the urban area) has been considered. Those budgets appear in fig. 4.9 as normally distributed with mean 1.43 h, and standard deviation 0.15 h.

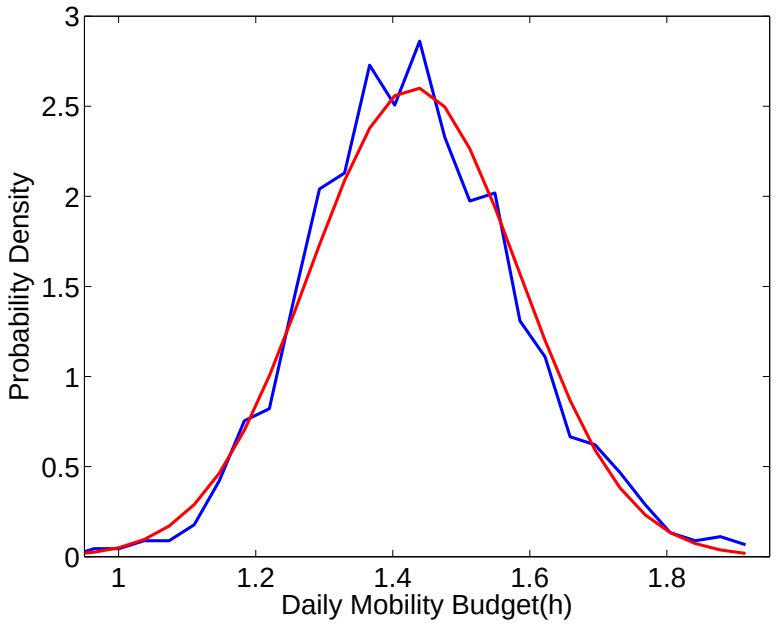


FIGURE 4.9: Distribution of the average travel time budget in 1233 italian cities. The red line represents the best-fit with a Gaussian distribution of mean 1.43h and std 0.15h.

Even if the values appear bigger than the expected 1.2 hours, the Gaussian interpolation and its relative small width is still suggesting a relative stability of the values of travel-time budget across cities of different size, economy an geographical characteristics. In order to better understand the differences between cities, we have picked the 29 of them where we have the larger number of drivers. This selection is a good sample of the whole set of cities, as their average travel time

budgets are spread in the same range (mean 1.40 h and standard deviation 0.12 h). For these cities, we have analyzed the shape of the travel time budget probability densities considering all the days of mobility of their inhabitants.

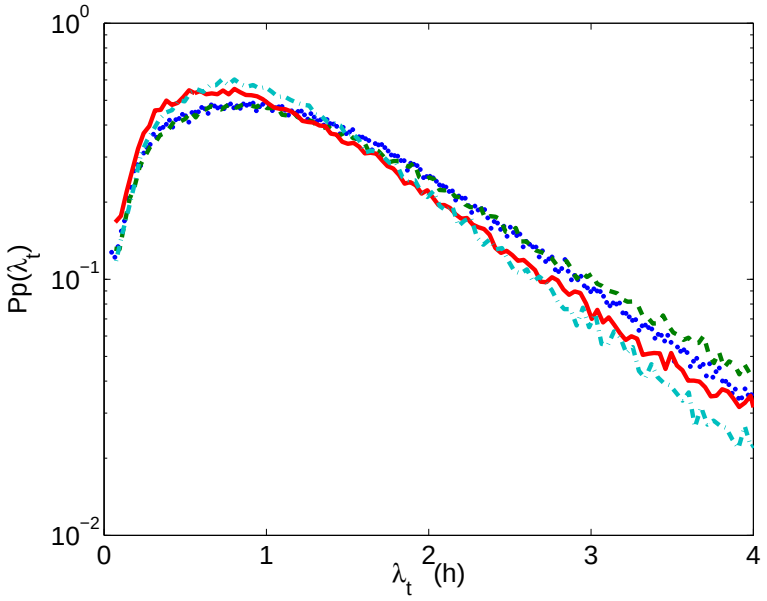


FIGURE 4.10: The travel time budget probability density for the cities of Rome (blue dot line), Naples(green dash line), Milan (red solid line) and Turin(cyan dot-dash line).

Those distributions are shaped similarly to fig. 4.5 right, as we can see comparing it with those relative to the 4 bigger cities in Italy that are in fig. 4.10. A notable phenomena we remark is that restricting the analysis to city dwellers the average travel time is bigger than what we find considering larger areas, in contrast with what suggested in [40]. In fact, limiting the analysis of the Emilia-Romagna dataset to only the city area of Bologna results in a growth from 70 to 89 minutes for the average travel-time budget. Some differences can be noticed between the values in Bologna in the two different years, but it can be the product of a slightly different filtering procedure.

Distributions similar to those of fig. 4.10 have been interpreted in [37] as a realization for a particular mode of transportation of an universal travel energy distribution (with the energy defined as $E_i = p_i t$ that is proportional to p, the energy consumption per unit time in the use of a particular transportation mean) and fitted with the curve:

$$p(t_b) \propto \exp\left(-\alpha/t_b - t_b/\beta\right) \tag{4.11}$$

The dominating term $\exp(-t_b/\beta)$ corresponds, similarly to what observed for the total mobility, to the entropy maximizing Maxwell-Boltzmann energy distribution. A second term, which in the reference paper[37] was chosen to be $\exp(-\alpha/t_b)$, describes the suppression of short trip. An interpretation for this suppression is that it is reflecting the fact that short trips are less likely to be undertaken, because there is an additional amount of energy for the preparation of a trip, of the order $\alpha p_i t$, that makes the energy budget not worth to be spent. We add to this interpretation, being our data relative to an unique transportation mean, the car, that this suppression might represent also day of mobility where other means have been chosen.

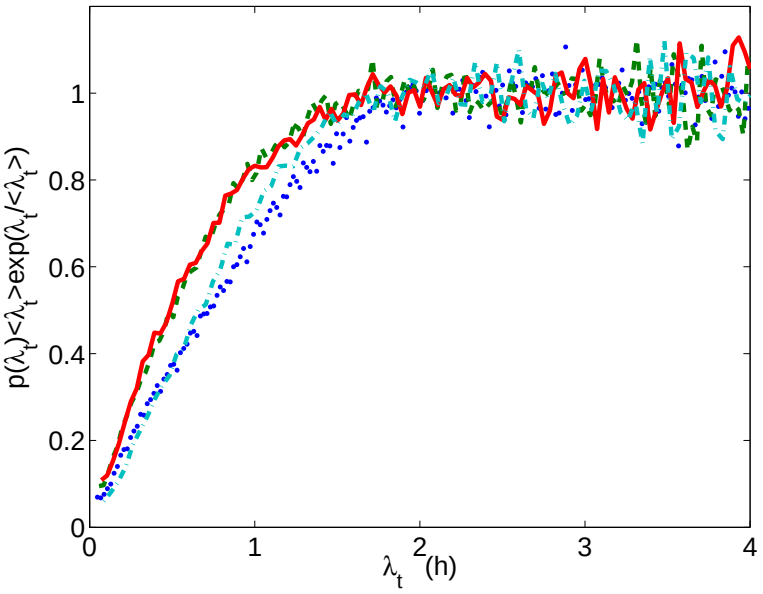


FIGURE 4.11: Suppression function for the cities of Rome (blue dot line), Naples(green dash line), Milan (red solid line) and Turin(cyan dot-dash line). This function is obtained dividing the probability densities of figure 4.10 for the exponential best-fit for the exponential tail.

In our data, the presence of an exponential tail is clearly suggesting the same dominant term. Multiplying the empirical distribution $p(t_b)$ by a best-fit for the exponential tail $\frac{1}{\beta}\exp(-t_b/\beta)$ we can isolate the suppression term function, that is represented in fig. 4.11. Many interpolating functions can be proposed but none appear in a manifest agreement with the observed suppression functions. Therefore, in order to make a comparative analysis of the trip suppression among different cities without relying on a particular interpolating function, we define a time t_c where the suppression function reaches the value 0.5: this value may represent an amount of time budget for which it becomes worth to take the car

for performing a part of the daily mobility. Comparing our approach with the $\exp(-\alpha/t_b)$ curve fit, we have a linear relationship $t_c = \frac{\alpha}{\ln 2}$.

The average value of travel time budget $\langle t_b \rangle$ grows when either β or t_c grow, stretching the tail of the distribution when incrementing β and excluding small values when incrementing t_c. This has been also observed across all cities with a correlation of 0.88 between $\langle t_b \rangle$ and β and a correlation of 0.55 between $\langle t_b \rangle$ and t_c. Noteworthy, the two quantities used for defining the distribution β and t_c have a correlation of only 0.15, granting that they are related to independent aspects of the drivers behavior. Measures of $\langle t_b \rangle$ hide these aspects, not considering the non Gaussian form of the distribution. For this reason, in the end of this chapter will focus on β or t_c, trying to isolate the key aspects shaping the total travel-time distribution.

For the 29 considered cities, we have gathered data on municipality surface and population from the Italian statistics institute ISTAT[43], on average income in provincial capitals, from data of the Ministry of Economy published by an economical newspaper[44] and on house prices (for the end of year 2011), taken from a vertical real estate search engine[45][6].

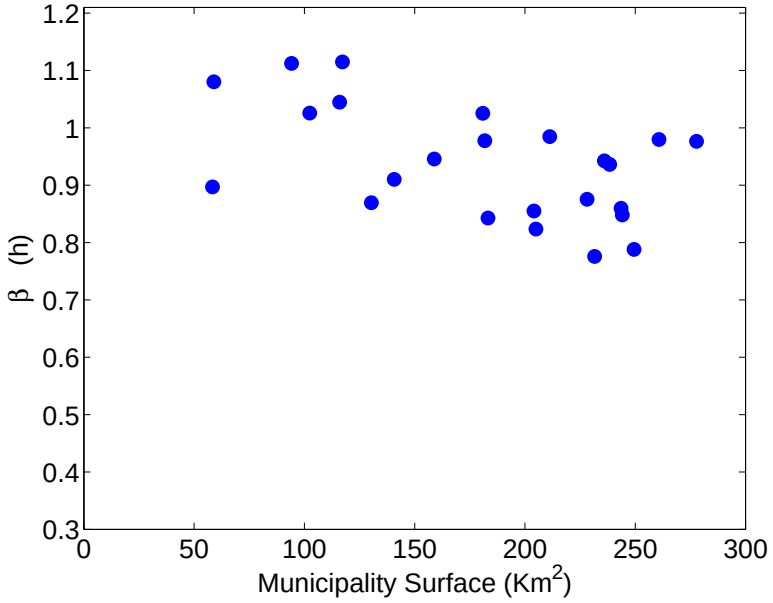


FIGURE 4.12: Relationship between municipality surface and β for cities which surface is under 300 Km^2.

First, if we isolate cities whose municipality surface is under 300 Km^2 we observe an anti-correlation of -0.56 between surface and β (see figure 4.12). This result is

[6]This study has been conducted with advises by Prof. Dirk Helbing.

somehow confirming what shown restricting the analysis from a region to a city: the values of the travel time budget tend to be more limited (β is a sort of energetic constraint) where the urban area is bigger. This fact can be a consequence of a higher level of stress, and thus higher energy consumption, when driving in bigger cities. At the same time, larger cities are offering more suitable activities one closer to another, so it can also be easier to optimize daily patterns. Excluded cities do not follow this relationship. But, those cities are Rome and others (Foggia, Perugia, Andria, and Grosseto) having a low population density. Probably, in both cases our value of surface is not a good measure for the size of the urban area where the mobility takes place because their municipality border include large parts of countryside (the municipality of Rome, for instance, includes many natural reserves).

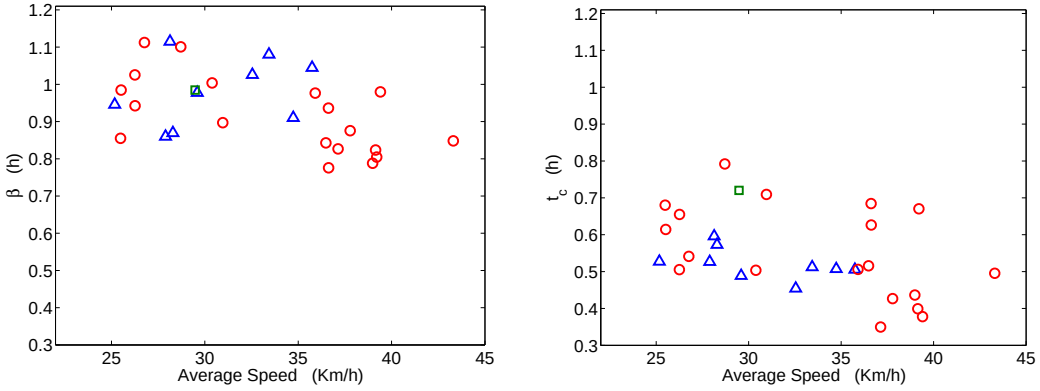


FIGURE 4.13: Relationship between average speed and respectively β (Left) and t_c (Right): red circles indicate cities with density under 2000 inh./Km^2, blue triangles cities over that threshold while Roma is represented by a green square.

Second, as we see in figure 4.13, both β and t_c are anti-correlated (with a coefficient of -0.48 and -0.44 respectively) with the average speed of traveling in the area, that we measure as the total distance covered in the month by all drivers divided by the total travel-time. It is clear that an high average speed is attained when there is an efficient road network. This efficiency lowers the activation energy that has to be overcome for choosing to perform the mobility (in particular to perform it by car) and therefore lowers t_c. It is even possible that optimized organizations of the street network that optimize of car travel time, e.g. an intense implementation of roundabouts, might turn the same network into something less accessible by other transportation means like bikes or walking and thus raising the activation energy for those other means. But is fundamental to remark the reduction of β when the average speed grows. In fact, this is evidently contrasting with the assumption of

a constant travel-time budget: if the street network is more efficient the activities that an individual wants to perform can be reached with less effort, but we see that travel-time savings due to a faster network are not invested in new mobility but are, indeed, saved.

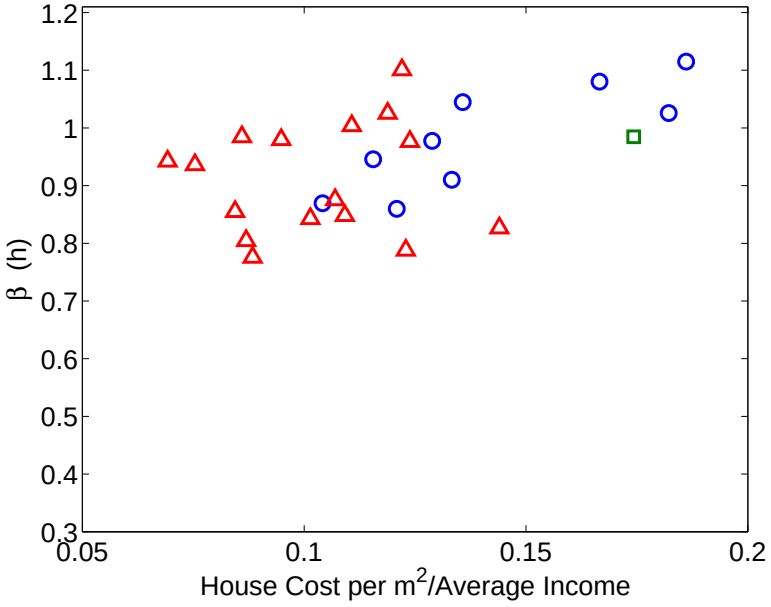


FIGURE 4.14: Relationship between relative housing price and β: red circles indicate cities with density under 2000 inh./Km^2, blue triangles cities over that threshold while Roma is represented by a green square.

In the analysis involving economical quantities, we have found that house prices is a key quantity related to both β and t_c. In the relationship with β, a good correlation (0.50) has been found with the house prices per square meter rescaled with the average income, a quantity that probably represents better the impact of housing on families' financial resources than only the house price. An even higher correlation of 0.78 is present for cities with a population density greater than 2000 inhabitants/km^2 (circles and squares in figure 4.14). This indicates that relative housing costs influence those major cities mobility, and do that more than for minor cities. This difference can be related to work-related migrations, which are more frequent toward big cities. In fact, in case of migration, higher relative housing costs leads to the choice of a living place further from the city center, where the most attractive activities take place. This tradeoff is consistent with the modeling framework of radiation model[36], assuming a not uniform benefit distribution $p(z)$ at a urban scale.

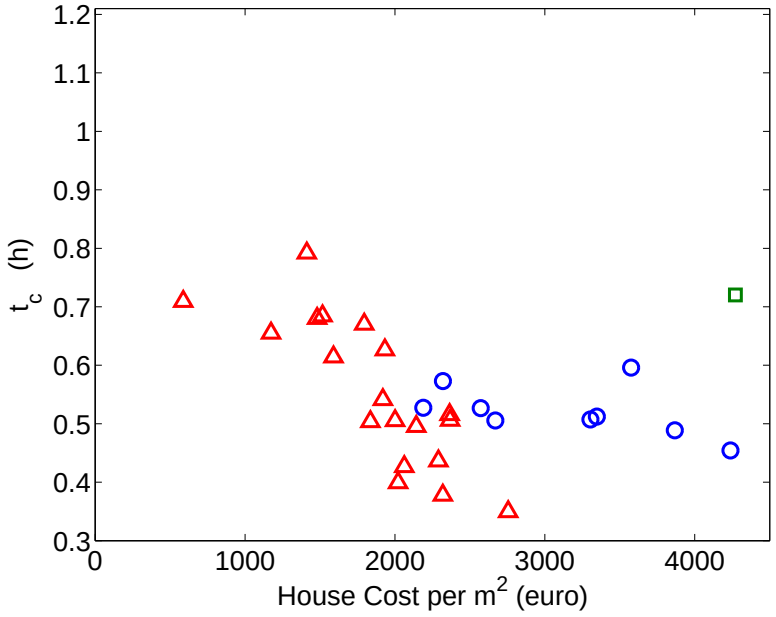


FIGURE 4.15: Relationship between housing price and t_c: red circles indicate cities with density under 2000 inh./Km^2, blue triangles cities over that threshold while Roma is represented by a green square.

In the last figure, house prices have been found anti-correlated with t_c. This is especially true (-0.81) for cities with a density under 2000 inh./km^2 (triangles in figure 4.15) where there is also a broader range of values for t_c. For denser cities, the activation time is instead almost constantly around the half hour: probably, under this threshold, the time needed for going to and finding a parking place overcomes the advantage of using the vehicle. This appears not true for less dense cities, where the anti-correlation with house prices suggests that the activation effort is influenced by economical factors. Both the correlation coefficients of t_c with the average income and with the relative house price are lower, so housing cost appears to be the better indicator for this influence. This anti-correlation can be related to a lesser impact of monetary costs related to the use of cars where economical conditions are better. If those costs are considered smaller with respect to the utilities associated with the performed activities and to time saving and higher comfort due to the use of a vehicle, the rational choice for a driver becomes using the car even for short travel-times.

In summary, our analyses show:

- travel-time budget is not constant, and can be described using two parameters: β and t_c;

- β is smaller, if the city area is great;
- both β and t_c are greater, if the road network is efficient;
- β is greater, if the relative costs for housing are high (this is especially true for major cities);
- t_c is smaller, if the costs for housing are high (this is especially true for minor cities).

Peculiarly, with her combination of extreme house prices and gigantic municipality size, Rome does not fit some of these schemes.

Chapter 5

Use of Space

In the precedent sections, we have discussed of the statistical properties of the trip lengths and durations with studies that have been intentionally developed not considering the environment where the mobility takes place. For a successful modeling of Human Mobility, it is necessary to understand not only how much people are willing to move, but also how these movement are distributed into space. The approach that is commonly used in transportation modeling for this purpose is the origin-destination matrix, which is obtained by a division of the of interest into different zones $i = 1, \ldots, N$: the number of individuals going from i to j defines the matrix T_{ij}. This matrix is equivalent to a directed and weighted network and describes the aggregated mobility of all drivers in a delimited area and a definite time span. This network is suitable for studying the rush hour mobility of commuters, which can be often deduced from census questionnaires. A model recalling Newton's law of gravity, and therefore called gravity law, is the prevailing tool to predict the origin-destination matrix. This law states that T_{ij} is proportional to the some power of the populations living in i and j, and decays with a function $f(r_{ij})$ of the distance between the two locations. Although, a novel method called radiation model[36] has been recently proposed as an replacement to gravity law. Still using only population densities as input quantities, this last model considers also the possible alternatives to j when choosing a target destination form i.

Aside from this aggregate point of view, thanks to the advances in Information and Communication Technologies, it is now possible to focus on individuals' mobility. The description of the spatial mobility of a particular person involves many different quantities like radius of gyration, confidence ellipses or Zipf exponents of the distribution of visited locations frequencies[21][14]. In the following, we focus on the characteristics of the individual mobility network that each person defines

with their movements from a location to another. In particular, we study the degree and rank distributions scaling exponents and how these are related to the way people explore new places.

5.1 Individual Mobility Networks

5.1.1 Degree Distribution

Each person's mobility differs from the others because of the own habits, agenda and knowledge of the urban environment. To study those differences, we introduce individual mobility networks, in which a node represents a visited location where an activity has been performed, while each weighted and directed link implies the existence of one or more trips between two locations[1]. Our objective here is to study the general features of individual mobility networks and in particular the hierarchy between performed activities, which is pointed out by the degree of a node. The degree quantifies the total number of connections or, more specifically for our case, the total number of trips that have passed by a node. As trips start from the same location where the last trip stopped, even if the network is directed in- and out-degree are equal and thus we will use only the term degree. Although, a single mobility network, derived from one month of mobility, is often not enough for an accurate statistical analysis. For this reason, we have extrapolated the shape of the individual networks' degree distribution by superposing the distributions of the whole ensemble of networks. This aggregated degree distribution is shown in figure 5.1, where we have highlighted its scale-free behavior with a power law fit.

[1]Being the weights the number of passages through one link, the network is more specifically a multigraph.

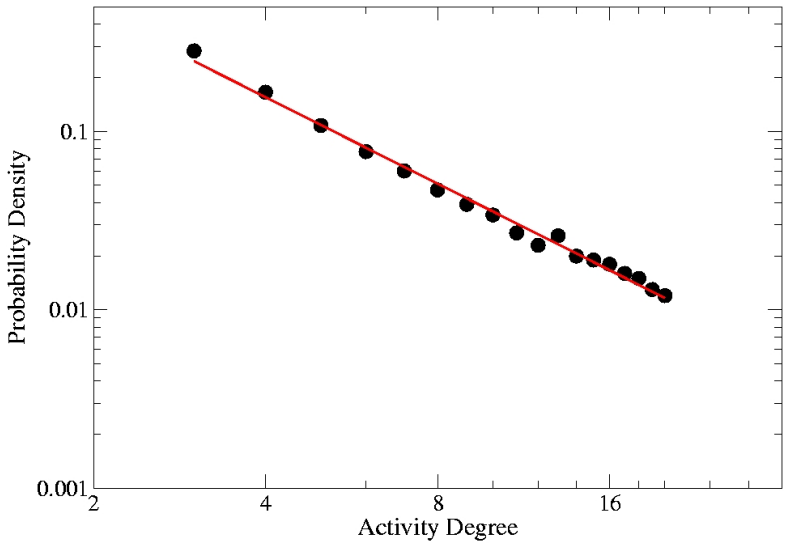


FIGURE 5.1: Empirical distributions of the activity degree (circles) in Florence. In a log-log scale we enlighten the interpolation with a power law k^{-b} with $b \simeq 1.6$ (continuous curve).

Dealing with a scale-free network, it appears natural to verify if it is possible to modeling it with preferential attachment[46]. The extreme simplicity of this model makes it highly adaptable to the description of a wide range of phenomena and for thus represents one of the pillars of network theory. In this sense, we have developed a Monte Carlo simulation where a weighted network expands with a succession of trips each starting from the last arrival node. If destinations are chosen with preferential attachment with weight proportional to connectivity, the scaling exponent found is close to -2[47]. With respect of the Barabási-Albert model, besides the constraint of following a path, here the number of possible note is in principle limited, as the number of places in the analyzed area are finite, and a "freedom" parameter has been introduced to permit the exploration of locations different from the two extremes of the first trip chosen as initial condition.

The difference of the scaling exponent observed in our data from -2 is due to the fact that working time schedule introduces further constrains to the individual mobility agenda. Activity times statistical properties have been included in a microscopic model where individual human mobility is represented as a dynamical process on a weighted network, in which each individual jumps from one node to another in a random way with preferential attachment.[17]. This model is successfully self consistent in explaining cell-phone users mobility. However, as we have already remarked in the chapter on time use, visit durations estimated with cell phones possess statistical characteristics inconsistent with our observed downtime durations, which are following the Benford's Law (eq. 3.1): the key

factor in this microscopic model is the deviation of the scaling exponent of activity durations from -1, that in our case is negligible. For this reason, we propose a different model to include activity duration following the Benford's Law.

We have proposed in section 3.2 the existence of a universal distribution $f(u)$ for the normalized activity downtime (eq. (3.4)). Now, the existence of an universal distribution implies (cfr. eq. (3.5)):

$$p(t,k) = f\left(\frac{t}{< t >_k}\right)\frac{kp(k)}{< t_k >} \tag{5.1}$$

Then using the interpolation (3.6) and performing the change of variable $u(k) = t/\exp(\gamma k^a)$, we obtain that the Benford's law implies a power law distribution for the activity degrees (see appendix C):

$$p(k) \simeq \frac{1}{k^{2-a}} \tag{5.2}$$

According to the estimate (3.6), we expect an exponent $\simeq -1.7$. In fig. 5.1, we plot the empirical activity degree distribution with a numerical interpolation by a power law k^{-b}; data provide $b \simeq 1.6$ which is consistent with the analytical estimate (5.2).

5.1.2 Heaps' Law of the Number of Visited Locations

Alternatively to degree distribution, in scale free networks also the rank distribution is suitable to describe hierarchy among nodes. Rank distribution can be obtained ordering nodes from the highest degree to the lowest one for each individual mobility network. A simple relationship links the scaling exponents of the two distributions[19]: being b the exponent for the degree distribution, the distribution of ranks r reads:

$$p(r) \propto r^{-\frac{1}{b-1}} \tag{5.3}$$

We examine the rank distribution grouping the individual mobility networks according to the number of nodes and computing the average visitation frequencies f_r for nodes with the same rank. The results are reported in the figure 5.2, where we point out a possible interpolation with a power law distribution $f_r \propto r^{-\alpha}$ where the exponent $\alpha = 1.42$ is in agreement with the analogous results on human mobility based on a different data set[17]. Due to the relationship between degree and rank, the existence of a power law distribution for the frequencies rank

indicates that the individual activities network is structured according to a preferential attachment rule, where the most visited locations could be related to habit mobility.

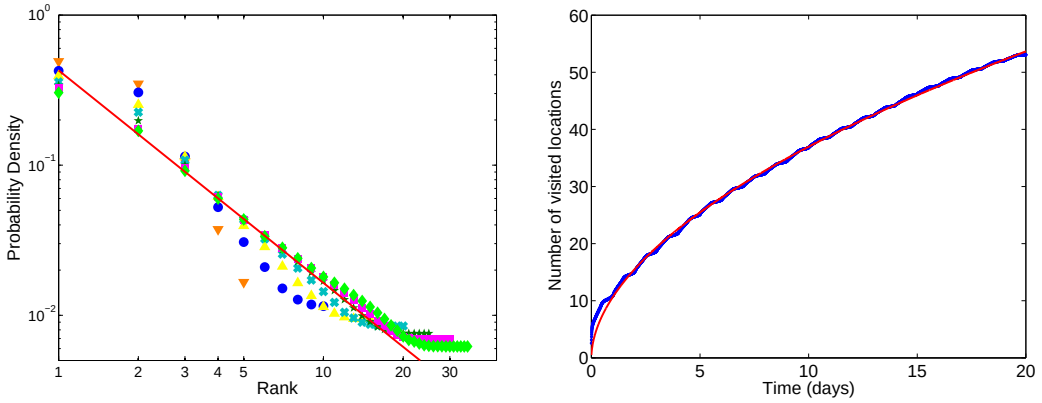


FIGURE 5.2: **(Left picture)** Rank distribution of average visitation frequencies for each nodes in the individual mobility networks with a fixed number of nodes: $N = 5$ (triangles down), $N = 10$ (circles), $N = 15$ (triangles up), $N = 20$ (light squares), $N = 25$ (stars), $N = 30$ (squares) and $N = 35$ (diamonds). The red line corresponds to a power law interpolation where $f_r \propto r^{-\alpha}$ where $\alpha = 1.42 \pm .06$. **(Right picture)** Number of visited distinct locations as a function of time (day unit); the continuous line is an interpolation using a power law $n(t) \propto t^{\beta}$ where $\beta = 0.5357 \pm 0.006$. In both figures the data refer to individuals that perform at least 20 mobility days in the Emilia-Romagna dataset.

A power law rank distribution is commonly known as Zipf's law, after the name of the linguist that first proposed a similar law for the frequency of any word[48]. Strongly related to Zipf's law is Heaps' law[49], stating that the number of different words used in a text grows as a power of the length of the text itself[50]. In our case, the equivalent of Heaps' law is the study of the diffusion process where, over time, the number of visited location grows. We have empirically studied this process using the time dependence of the total number of different visited locations $n(t)$ for individuals whose mobility covers at least 20 days in the analyzed period: i.e. $n(t)$ is the number of new locations visited by the ensemble of individuals in a time t. We apply a Markov hypothesis to describe the evolution of $n(t)$. Letting $p(r,t)$ be the probability that individuals have visited r locations after t days, we have the Master equation

$$p(r,t+1) = p(r,t)\bar{p}_r + p(r-1,t)(1-\bar{p}_{r-1}) \tag{5.4}$$

where $\bar{p}_r$ is the average probability of choosing one of the r visited locations, that we assume not dependent from t (stationary process). By definition:

$$n(t) = \sum_r r p(r,t)$$

so that from the equation (5.4) we get:

$$n(t+1) = n(t) + 1 - \sum_r p(r,t)\bar{p}_r \tag{5.5}$$

where we have used the normalizing condition:

$$\sum_r p(r,t) = 1$$

and we have neglected the boundary effect of a finite number of locations. To proceed, we need to estimate $\bar{p}_r$. Assuming that individuals perform a Markov's dynamics, $\bar{p}_r$ is the measure of the r visited locations. The average visitation frequency f_r can be interpreted either as a measure or as a choice probability of the r location. Let us order the r locations according to their rank, the average measure of the $j \in [1,r]$ choice (after $j-1$ choices), m_j, can then be estimated with:

$$m_j \propto \int_j^N f_l^2 dl \propto \int_j^N \frac{1}{l^{2\alpha}} dl \propto \frac{1}{j^{2\alpha-1}} \qquad N \gg 1 \tag{5.6}$$

where we have used the power law interpolation for the rank distribution $f_r \propto r^{-\alpha}$. As a consequence, the expected measure for the r visited locations reads:

$$\bar{p}_r \propto \sum_{j=1}^{r} \frac{1}{j^{2\alpha-1}} \simeq \int_1^r \frac{1}{j^{2\alpha-1}} dj \propto \left(1 - \frac{1}{r^{2\alpha-2}}\right) \tag{5.7}$$

By definition, $\bar{p}_r \to 1$ as r increases. Using the estimate (5.7), the equation (5.5) reads:

$$n(t+1) = n(t) + 1 - \sum_r p(r,t)\left(1 - \frac{1}{r^{2\alpha-2}}\right) = n(t) - \sum_r \frac{p(r,t)}{r^{2\alpha-2}} \tag{5.8}$$

Then, we apply the mean field theory argument to reduce the equation (5.8) to the simple form:

$$n(t+1) = n(t) + \frac{1}{n(t)^{2\alpha-2}} \tag{5.9}$$

whose solution can be approximated by:

$$n(t) \simeq ct^{1/(2\alpha-1)} \tag{5.10}$$

where c is an integration constant[2]. According to the f_r interpolation (fig. 5.2 left) $\alpha \simeq 1.42 \pm .06$ and we get:

$$n(t) \propto t^{\beta}$$

where $\beta = .54 \pm .03$. This result is very close to the numerical interpolation of the empirical measures $\beta = .53$ (fig. 5.2 right) and it has to be compared with the result ($\beta = .60 \pm .02$) and the relative individual mobility model developed on mobile phone data[17], the model we said that cannot be applied in our case since it is not consistent with the empirical activity time distribution. The results may be interpreted in a twofold way. From one hand this is another indication that macroscopic statistical properties of human mobility mimics the properties of an ensemble of particles which perform a stochastic Markov dynamics, taking into account the existence of spatial and temporal constraints. On the other hand the individual dynamics is certainly not a Markov process and the rank distribution in fig. 5.2 is the result of a cognitive behavior defining the daily mobility agenda in a complex urban environment.

[2]More generally, equation 5.10 can in principle be extended to any Zipf 's law with $\alpha > 1$.

5.1.3 Mobility Classification: Mono and Dipolar Networks

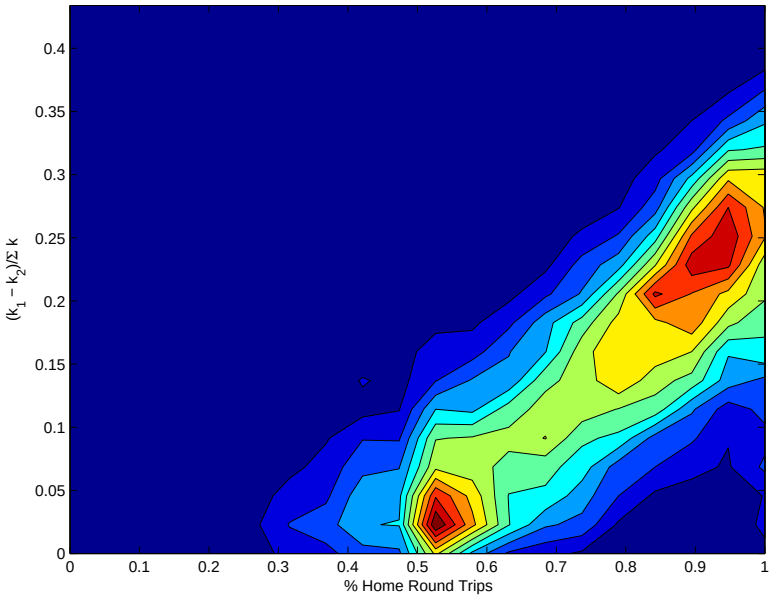


FIGURE 5.3: 2D distribution of the percentage of round trips having at one hand the house vs. the relative difference of the two main hubs degrees (Florence Dataset). We observe that the distribution is manifestly bimodal, where the two peaks represent two different spatial organizations of trips, one dipolar and another mono-polar.

One difficult task that we have undertaken is to classify individual mobility network. Various network theory quantities have been evaluated (see Appendix D) in order to find characteristics allowing to separate the networks in different categories. However, all the considered distribution are bell-shaped, and thus do not allow to define classes of networks. On exception is represented by the difference between the degrees of the two main hubs. These hubs are the nodes with the higher number of connections and therefore represent the most important locations of the mobility network. If we compute the difference of the degrees k_1 and k_2 of the first and second most important hubs (that can be seen, for instance, as home and workplace), rescaled by the sum of the degree of the whole network $\sum_i k_i$ in order to be able to confront network with a different number of trips and nodes, we find a distribution that recalls the superposition of two bell shaped functions. In order to permit a better identification of this bi-modality, we can use a second quantity that we can extract from the mobility pattern of visit of the network: the number of round trips that have at one end the main hub, namely the home. If we represent these two quantities in a 2D distribution (fig. 5.3), we can clearly distinguish two peaks. The one at the bottom of the figure represents

what we may call "dipolar networks" where two almost equivalent hubs exists ($k_1 \simeq k_2$), and only half of the round trips have the house at one end. The one on the right of the figure represents "mono-polar networks' where the second hub is significantly less visited than the main hub, and almost all the round trips have the main hub at one end. These two classes of networks embody two different spatial organizations of trips and probably two different way of interacting with the urban environment. From the empirical distribution, we can deduce that the majority of the networks are mono-polar, but further analysis have to be done to understand causes and impacts of these two types of mobility. It is possible that other classes, as tri-polar networks, exist, but are so infrequent that they are not distinguishable it the considered distribution.

5.2 Interaction Network

In the second part of this chapter we move from individual to collectivity, analyzing how places attract people and how people share places. This analysis is defined studying interactions networks that we define as a bipartite graph[3] where the nodes are either individuals or locations and a connection exists if a location has been visited at least once by that individual. From this bipartite graph, two different networks can then be defined. One is the individual interaction network, where individuals that have shared the same location are put in relationship, and the other is the location interaction network, where a link between two locations identify a driver who has visited at least once both places. In our opinion, these networks might be be useful to extrapolate information on social or simple proximity interactions eventually related the epidemic spreading. In this section, we propose a study on the statistical properties of the interaction network that has been obtained from the multi-driver clustering of the destination coordinates in the Emilia-Romagna dataset.

[3] A graph is bipartite if it can be divided in two subset of vertices such that no vertices in the same set are connected.

5.2.1 Attraction Basins

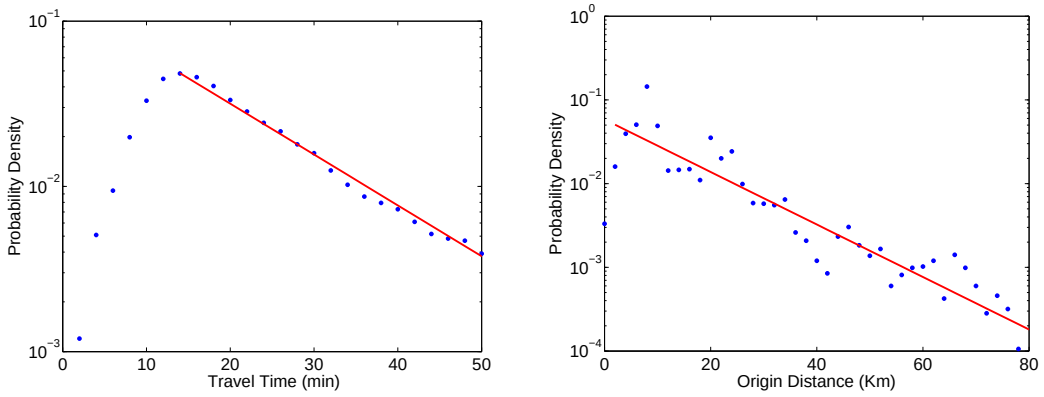


FIGURE 5.4: **(Left picture)** Distribution of travel durations to the considered mall. The red line suggests an exponential interpolation for the tail. **(Right picture)** Distances of the origin of the trip to the mall, also here the red line suggests an exponential tail. Both graph are obtained from data of the Italy dataset.

Before dealing with the statistical properties of the Emilia-Romagna interaction network, it is important to have an idea of the attraction basin that a location can have. We propose here briefly the results of case study based on $\approx$ 8000 trips toward a mall situated in southern Italy, along a trunk road outside from any urban area. Analyzing both the distance of origin of costumers and the time they take for getting to the shopping center, we find that the attraction is short ranged. The probability of finding a trip that took a travel-time t follows a curve recalling the distribution of travel-time budgets (fig. 4.10), with a exponential tail (here with a scale parameter of 14 ± 1 minutes) and a suppression for brief trips under 10 minutes. Besides, the probability of finding a costumer at a distance d decreases also exponentially, with a scale parameter of 14 ± 2 Km and a less noticeable, but still present, suppression effect for close origins. We expected that close origins were rare, being the location situated far from any city. The similarities of these distributions with those of daily budgets make easy to believe that energetic traveling limits and effort-advantage balance can be fundamental also for the single choice among leisure activities.

5.2.2 Relation with Settlement's Fractal Dimension

The degree of a node i in the bipartite interaction network is the number I_i of different individuals who have visited at least once that node. As we can see

in figure 5.5, the probability of finding an activity visited by I_i individuals is distributed in the analyzed region following a power law $p(I) \propto I^{-\gamma}$ with an exponent $\gamma = 1.46 \pm 0.04$.

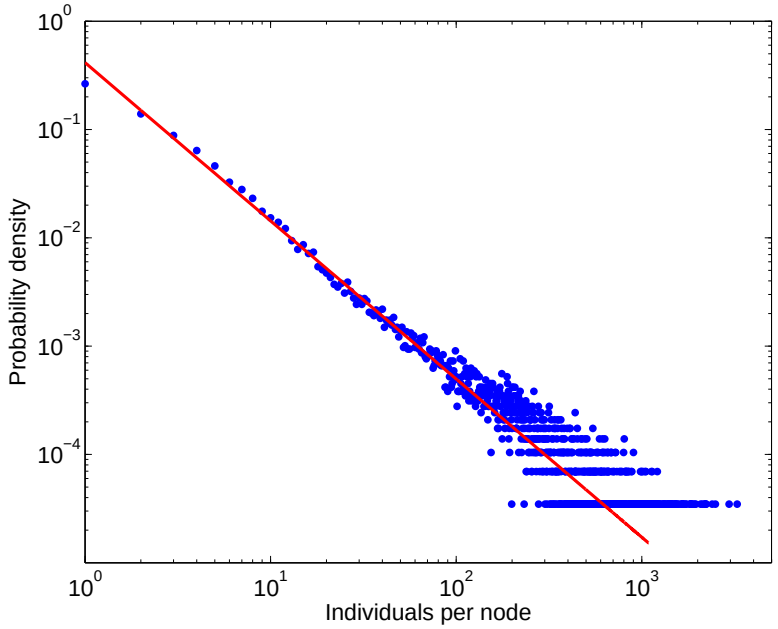


FIGURE 5.5: Distribution of the number of different individuals per node: the red line represents the best-fit with a power law $p(l) \propto l^{-1.46}$, calculated in the range 1-100.

In the last section we found evidence that the spatial attraction basin of a particular location can be reasonably modeled by an exponential radial distribution with a scale parameter R. Let us assume that this attraction distance is equivalent to an "excluded area" $\sigma(R)$, representing the attraction basin of the considered location. If we assume that the attraction area is the same for all similar activities, the total number of conflicting activities A with attraction basin R that can be present in a region of given surface is:

$$A \propto R^{-2}$$

Being the region characterized by a fractal geography of the lived areas, and if we assume a constant population density in such area, the number of inhabitants that can be found in the attraction basin of radius R is:

$$I_\sigma \propto R^D$$

Where D is the fractal dimension. Therefore:

$$A \propto I^{-\frac{2}{D}}$$

meaning that the number of activities that can be found with I inhabitants in its attraction basin scales with a function of the fractal dimension. Superposing the effect of the conflicting area for all the different activities and locations we find in general:

$$p(I) \propto I^{-\frac{2}{D}} \tag{5.11}$$

This relationship is confirmed by our experimental analysis on Emilia-Romagna, where the fractal dimension found for dimension between 500 and 600 meters (the clustering procedure is defining areas of at least 500 m of radius) is $D_{er} = 1.31 \pm 0.03$, while the dimension that would justify the scaling law of individual presence is consistently $D^*_{er} = 2/\gamma_{er} = 1.37 \pm 0.04$. An analogue result has been found dividing the Florence dataset in squares with edge 500 m: in this case the $\gamma_{fi} = 1.4 \pm 0.1$ while the fractal dimension is $D_{fi} = 1.4 \pm 0.1$.

5.2.3 Location Interaction Network

The location interaction network is represented by an undirected multigraph[4] where nodes represent locations and each link represents an individual who have visited the two locations. This kind of network is representing an intermediate point of view between Origin Destination matrices and individual mobility networks: the location-location structure may recall the network derived from the Origin-Destination matrix, but each link here contains information not of direct movements but of the whole mobility of the driver. The degree of a node in this multigraph is given by the sum of the connections due to individuals who have visited the same node i. Each individual participates to this sum with the N_p nodes present in his individual mobility network:

$$d_i = \sum_{p=1}^{I_i} N_p \tag{5.12}$$

Where sum has to be intended over the I_i individuals labeled by p who have visited the node i.

[4] A multigraph is a graph in which are permitted more than an edge between two vertices.

From the empirical analysis, it emerges that between d_i and I_i there is a linear relationship with $\frac{d_i}{I_i} \approx 30$ at all scales (see figure 5.6).

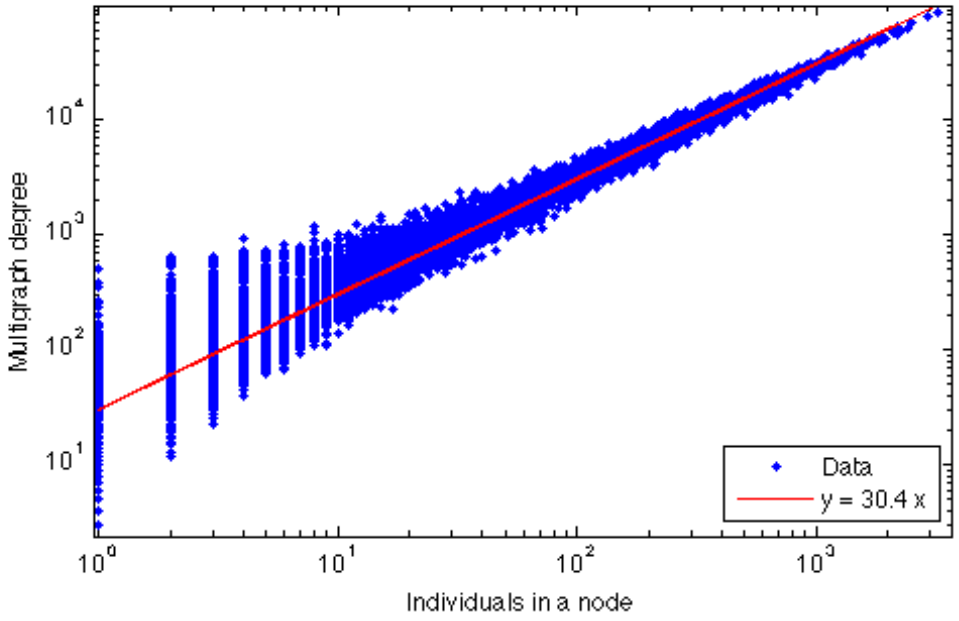


FIGURE 5.6: Linear relationship between number of individuals and multigraph degree of the nodes. The superposition with the red straight line reflects the validity of eq. 5.17.

This relationship is due to the fact that the probability of finding an individual having N_p connections in a particular node i of the network is proportional to N_p, as individual with bigger mobility networks have been present in a larger number of nodes, multiplied by the probability of having a individual that have visited N_p nodes in our data:

$$p(N_p) = C_{norm} \cdot N_p \cdot \frac{1}{\sqrt{2\pi}\sigma_N N_p} \exp\left(-\frac{(\log N_p - \mu_N)^2}{2\sigma_N^2}\right) \tag{5.13}$$

where μ_N and σ_N are those found in Appendix E.2 for the empirical lognormal distribution of the number of nodes visited in the different individual mobility networks:

$$p(N) = \frac{1}{\sqrt{2\pi}\sigma_N N} \exp\left(-\frac{(\log N - \mu_N)^2}{2\sigma_N^2}\right) \tag{5.14}$$

Thus, the normalization factor C_{norm} is given by the inverse of the average number of nodes visited by drivers:

$$C_{norm} = \frac{1}{\langle N \rangle} \tag{5.15}$$

And the average value of the distribution $p(N_p)$ is the ratio between the second moment and the first moment of the distribution 5.14, ratio that has been numerically evaluated as 30.4:

$$\langle N_p \rangle = \frac{\langle N^2 \rangle}{\langle N \rangle} \tag{5.16}$$

Therefore, it is valid the relationship:

$$d_i = \sum_{p=1}^{I_i} N_p \approx \langle N_p \rangle \cdot I_i = \frac{\langle N^2 \rangle}{\langle N \rangle} I_i \tag{5.17}$$

binding the statistical properties of location interaction network's degrees to those of the distribution of individuals and those of the number of nodes in the individual networks.

Moreover, the weights in the network may result a fundamental quantity: for instance, knowing what are the strongest links in this network permit to identify key routes of disease transmission. Therefore, we conclude this chapter examining the values in the adjacency matrix A_{ij} of the location interaction network. This matrix is by nature symmetric, and diagonal elements A_{ii}, topologically representing loops, are exactly the number of individuals present in the node labeled with i. Network characterization is therefore completely covered by the strictly upper triangular part of the matrix. Evaluating in an histogram the values of A_{ij} in this area, we find that they are distributes following a power law with exponent close to -2.

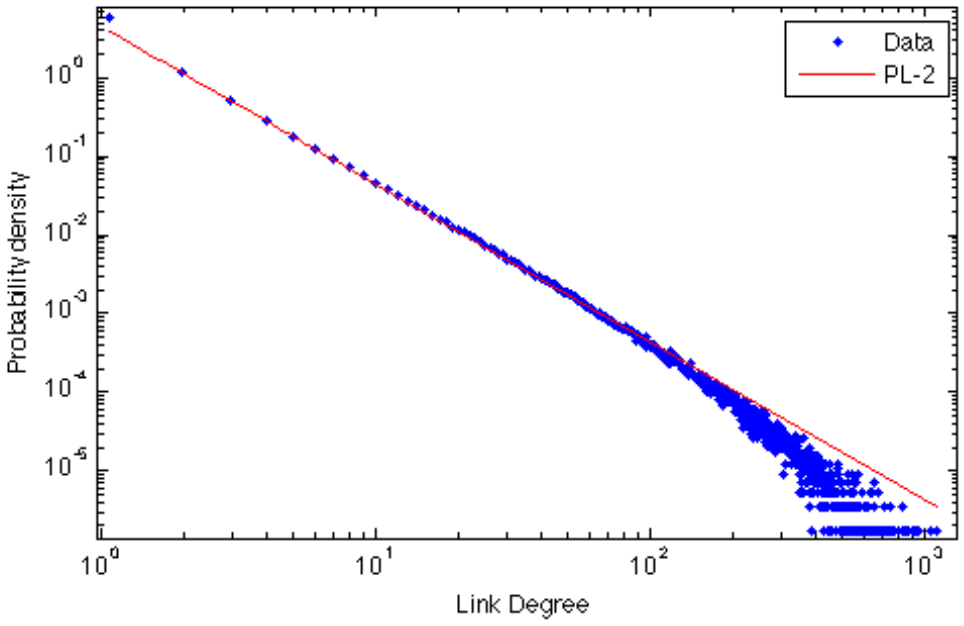


FIGURE 5.7: Probability density for the values of weights in the location interaction network: the red line suggests an interpolation with a power law of exponent -2.

Again, also this statistical property is completely determined by the individual visit distribution over the various locations. In fact, if individuals are distributes following a power law with scaling exponent ≈ -1.5, the related rank distribution has, after equation 5.3, a scaling exponent of ≈ -2. Therefore, we define:

$$p(i) = C_n i^{-2} \tag{5.18}$$

representing the probability that, extracted a random individual, they participate to node i. Given the relative weights to two nodes $w(i)$ and $w(j)$, let us suppose that the weight of a link $w(i,j)$ is proportional to the product $w(i)w(j)$. From this, we obtain the distribution $w(i,j)$ passing to the continuous limit from the given distributions $w(i)$ and $w(j)$ and with the coordinate change $x = i, y = ij$

$$\begin{aligned} &p(i)p(j)didj = \\ &p(x)p(\frac{y}{x})dxdy\frac{1}{x} = \\ &C_n^2 x^{-2}(\frac{y}{x})^{-2}\frac{1}{x}dxdy = \\ &C_n^2 y^{-2}x^{-1}dxdy \end{aligned} \tag{5.19}$$

Integrating over x we obtain $p(i,j) \propto y^{-2}dy$ and therefore the distribution of the link degree is a power law with exponent -2.

In closing, the results of this section indicate that the scaling properties of the location interaction network, i.e. way people use and share space, is fully determined by the characteristics of the territory, embodied by the fractal dimension. In defining the interaction network, we have excluded all information on the structure of the individual mobility network, and therefore ignored individual dynamics represented by the individual mobility networks. For a deeper understanding of individuals dynamics, in the next section we propose a study of the their mobility patterns where we take advantage of information entropy measures.

Chapter 6

Entropic Analysis of Mobility Patterns

At the beginning of every new day, people wake up knowing that a series of tasks have to be carried out. Necessarily, they must satisfy the physiological need of eating, that represents a periodical constraint in the daily routine, and sleeping, that forces the circadian rhythm and thus imposes an end to the chain of performed activities. Besides, they have to perform duties, usually the working related ones, that are precisely scheduled and others, as shopping, social and leisure activities, which can be done in any moment during the free time. Changes of plan during the day are always possible: a planned meeting may be postponed, a new activity may be picked instead of another or something can be done just because some spare time is available.

The accomplishment of the daily duties is realized traveling through the locations where each particular task has to be performed, and is clearly the primary cause of human mobility. Therefore, understanding the structure of the daily activity pattern is a crucial step in the development of a model for human mobility.

Thanks to the commercial spreading of the Information and Communication Technologies, important steps have been made toward the characterization of individual mobility patterns: approaches in this direction have been made analyzing mobile phone [14][17][51][52][53], geographic online social networks [54] and private cars' GPS [21][15] datasets.

The aim of this section is to highlight the different roles that activities of different duration have in the structure of the individual mobility patterns. For doing

this, we have chosen to use entropy as the key quantity for our study and we have introduced methods of analysis different form those used in [51]. For our purposes entropy is a good observable because it resumes the information present in a sequence of characters (given by the statistical properties and the correlations due to the personal habits) in a real number. In particular, as the entropy per character in principle does not depend on the length of the sequence, we can then easily compare patterns of different length representing the mobility of different individuals[1].

6.1 Mobility Patterns

For the purposes of this analysis, we assume that every time the engine was off for more than 5 minutes, this stop can be associated to an activity performed near the parking place. Once the locations are identified, at each activity is associated an identification number that, in analogy with the studies of the entropy of texts, we call *character*. Mobility patterns are series of those characters representing individual mobility.

We have isolated, in the Florence dataset, two types of mobility patterns for each individual. The first is the time pattern, where the month has been divided in equal time intervals of length Δt_{tp} and at each time interval we have associated an activity (in this case, traveling is considered as an extra activity and therefore is identified by a specific character). As it is possible that many activities have been performed in the same timeframe, one activity has been randomly chosen in case of conflict regardless the activities relative duration. We make this choice, that is the same made in [51], in order to make our method consistent with this precedent study. As a consequence of that, for time patterns with time interval Δt_{tp}, activities shorter than Δt_{tp} might disappear from the analysis. The second type of pattern is the jump pattern, which consists in the sequence of visited locations. Here, we have introduced a time interval Δt_{jp} that represents the minimum time interval considered. Activities shorter than Δt_{jp} are excluded from the sequences. If, after this exclusion, two or more identical characters appear one aside the other, they collapse in an unique copy of the same character, so any repetition of the same character is neglected. Clearly, these two types of patterns carry different information. On one hand, the time patterns focus both the way the people moves

[1]This study has been conducted with advises by Prof. Mirko Degli Esposti.

in both space and time. On the other hand, the jump patterns focus only on the spatial relations between the visited locations.

6.2 Entropy Measures

Here, we consider three different measures for the information entropy of a mobility pattern: the actual entropy S, the temporal-uncorrelated entropy $S_u = -\sum_{j=1}^{N} p_j \log_2 p_j$ and the random entropy $S_r = \log_2 N$, where p_j is the probability of finding the character j in our sequence and N the number of unique characters in the sequence. S_r represents a maximal value for the entropy, because for representing each value of the sequence it is necessary at least a number of characters equal to the total number of visited activities. S_u is a better measure of the entropy, as it considers also the uneven frequencies of characters in the sequence, while still ignoring the possible compression due to the relative position of the activities. These three values are clearly bound by the relationship $S \leq S_u \leq S_r$.

The measures of the entropy per character S have been evaluated using the Lempel-Ziv algorithm estimator. This algorithm searches for repeated sequences of characters that may be exploited for the pattern data-compression. For a sequence of length n, the estimated value of entropy is:

$$S = \left(\frac{\sum_{i=2}^{n} l_i}{n} \right)^{-1} \ln n \tag{6.1}$$

where l_i is the length of the shortest string starting at position i that does not appear in the part of sequence up to position $i-1$ (included). The goodness of this estimate raises with the length n of the sequences and decreases with the broadening of the size of the alphabet N.

The more informative is a sequence, the more difficult is to predict how it may continue. The predictability Π is defined as the rate of correct predictions about the value of a character in the sequence knowing all the precedent characters. Low entropy is related to an high predictability and vice-versa. An upper bound Π^m to the predictability of a sequence can be computed as a function of the entropy S by the inversion of the formula:

$$-\Pi^m \log_2 \Pi^m - (1-\Pi^m) \log_2 (1-\Pi^m) + (1-\Pi^m) \log_2 (N-1) = S \tag{6.2}$$

This last equation is a consequence of the Fano's inequality[51] and the inversion is possible as long as N and Π^m are not too small.

6.3 Time Regularity

We have considered 28 days in the analyzed month excluding two public holidays and a Sunday and compensating when the daylight saving time has been introduced. For each day of the week we have 4 daily patterns and we have counted how many times the drivers were, in a given timeframe i of width 5 minutes centered on the instant t_i, in their most visited location. The so found average values of R (probability of finding in a given hour the user in his most visited location during that hour) across all timeframes for the different users is distributed within the $79 \pm 9\%$ confidence interval, at the margin of which lies the result regarding mobile phone data[51]. However, considering only the most visited location does not permit to take into account the complete information that lies in all locations visited in a given hour i. Also in this case, an entropic approach can here grant a most comprehensive analysis, because the information regarding all visited locations in a given timeframe can be taken into account. Thus, we have calculated for each i the regularity entropy: $S^R(i) = -\sum_{j=1}^{N} p_j(i) \log_2 p_j(i)$, where $p_j(i)$ is the probability for the driver to be found at the position j at the timeframe i. The average values of $S^R(i)$ across all users for the different days shown in fig. 6.1, where can be observed how every day have a peak of dispersion in the afternoon and has its minimum value in the late night. The working days appear substantially equivalent, if we exclude a growing tendency to spend time in unexpected places in the evening (and in Friday afternoon). Saturday is the more variable day while Sunday appears clearly to have a late beginning.

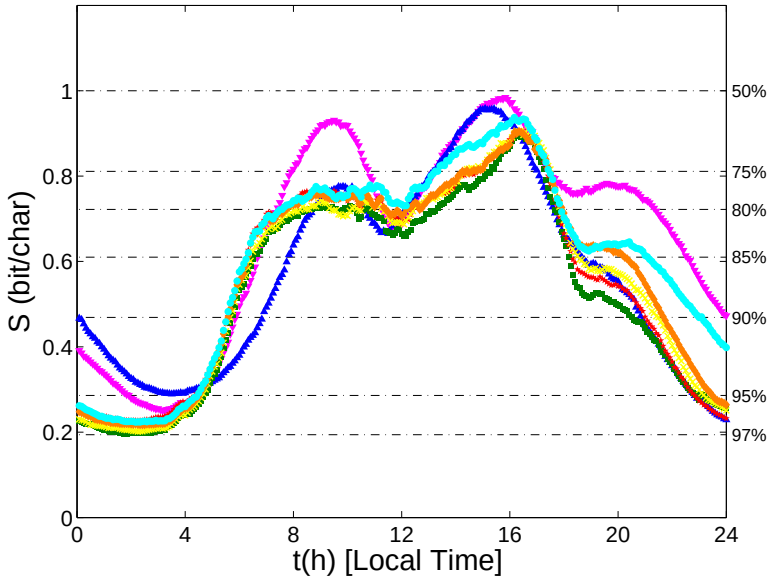


FIGURE 6.1: The average Regularity Entropy S^R measures the information lying in the distribution of the probabilities of finding a person in a particular place in a given timeframe. Every day of the week has a characteristic hour-dependency. Monday: green squares; Tuesday: red stars; Wednesday: yellow crosses; Thursday: orange diamonds; Friday: cyan circles; Saturday: magenta down triangles; Sunday: blue up triangles. With equation 6.2 we can link this entropic measure with the upper bound to predictability. The dot-dash lines represent the related values of $\Pi^m(S^R)$, calculated for $N = 2$.

6.4 Pattern Analysis

6.4.1 Time Patterns

Being our data extremely precise in time, we can afford to create time-patterns with relatively short values Δt_{tp}. This extension has a clear importance, because the downtimes in different daily activities follow a Benford's Law, i.e. a power law distribution with exponent near -1 (see chapter 3).

Therefore, we proceed examining the dependence of the entropy with respect of Δt_{tp}, analyzing mobility patterns created using different timeframes.

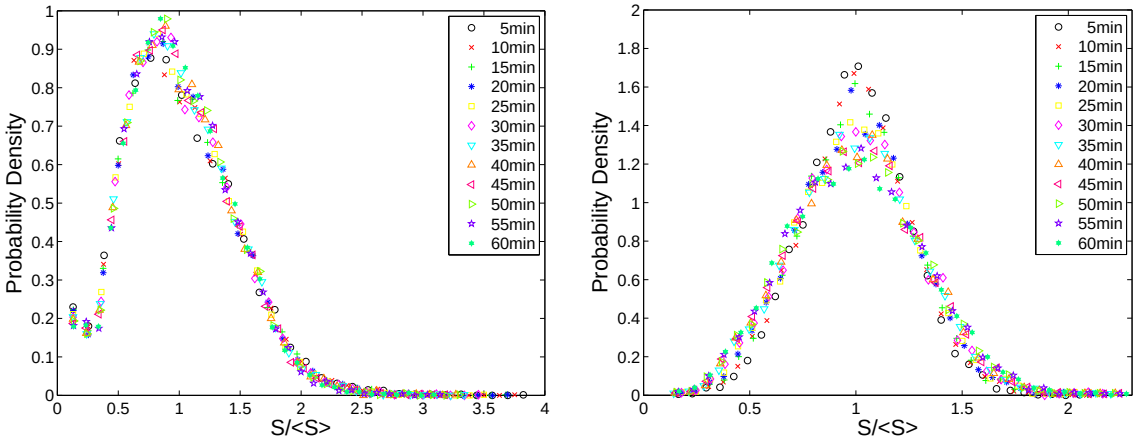


FIGURE 6.2: **(Left picture)** The rescaled distributions of the Time Patterns Entropies S with different timeframes Δt_{tp} are perfectly superposed, suggesting that all the dependency upon Δt_{tp} lies in the average value $\langle S \rangle$. **(Right picture)** Similarly, the rescaled distributions of the Jump Patterns Entropies S with different threshold Δt_{tj} are superposed and also in this case all the dependency upon the threshold lies in the average values of the entropy.

The results, regarding Δt_{tp} ranging from 5 to 60 minutes, suggest the existence of an universal probability distribution of the entropies S among the sample. In fact, the distribution of the rescaled values $p(S/\langle S \rangle)$ does not depend upon Δt_{tp}(fig. 6.2 left). Therefore, the only significant value representing the dependence of the distribution $p(S)$ on Δt_{tp} is the mean value $\langle S \rangle$. The same thing has been observed also for $p(S_u)$ and $p(S_r)$ and for the jump patterns. Thus, it is sufficient to study this dependence of the mean values of the entropy (see fig. 6.3).

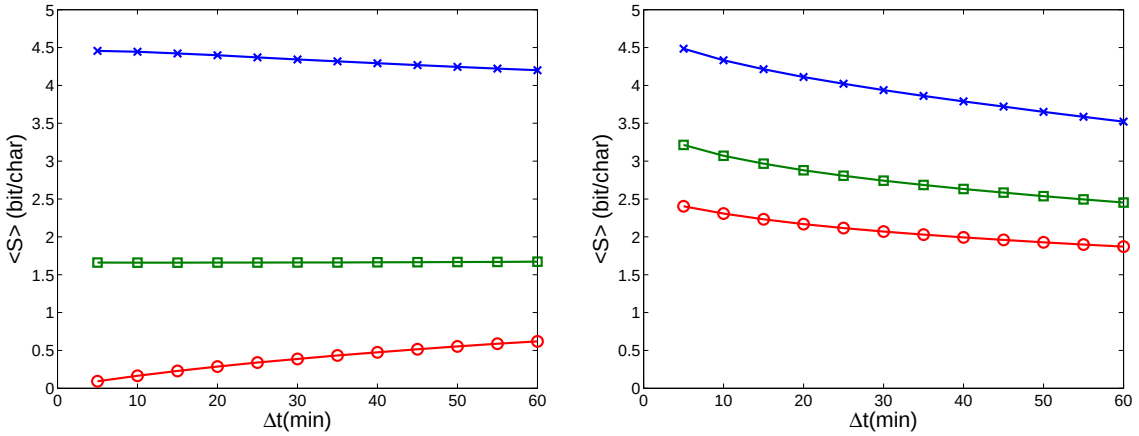


FIGURE 6.3: **(Left picture)**The dependency upon the timeframe size Δt_{tp} of the average values of the Time Patterns Entropies. $\langle S_r \rangle$ (blu crosses) grows for small values of Δt_{tp}, due to the inclusion of new locations, while $\langle S_u \rangle$ (green squares) is constant, indicating that the frequency distribution of the different characters in the sequences p_j is not dependent on the size of the timeframe. $\langle S \rangle$ (red circles) drops for small Δt_{tp}. **(Right picture)**The dependency upon the threshold Δt_{tp} of the average values of the jump Patterns Entropies. All the entropy averages $\langle S_r \rangle$ (blu crosses), $\langle S_u \rangle$ (green squares) and $\langle S \rangle$ (red circles) grows for small values of Δt_{tp}, showing that information increases when characters representing shorter stops are included.

The value found for $\Delta t_{tp} = 60$ minutes represents the same experimental conditions of the work of Song et al. on a mobile phone users' mobility[51]. Their results show an upper limit to the predictability of 93% that is in a remarkable accord with value of $\Pi^m = (92 \pm 3)\%$ that we can derive from our measures of S. The validation of this result on an independent, and rather different, dataset grants that mobility patterns obtained from both phone calls and private cars' parking are a good representation of human mobility with a temporal scale of one hour. But, by using a timeframe of this size, we are keeping out a great part of the mobility, as the 41% of the stops are shorter than one hour and thus are not considered in a 24-hour a day time pattern.

Going deeper in the analysis for shorter timeframes, we may observe that when Δt_{tp} decreases, S_r grows, S_u is constant and S drops, reaching values near zero. The growth of S_r is due to the growth of the number of different character in the pattern, as activities previously not observed are included when we, shortening the timeframe, create longer time patterns. The fact of S_u being constant means that frequency distribution of the characters p_j does not depend upon Δt_{tp}. This fact helps us in interpreting the tendency of entropy S toward zero for short timeframes. Indeed, this tendency suggests that the most part of the new characters introduced shortening the timeframes are only prolonging repetitions of the same character

representing long stops. Even if we are introducing new information regarding shorter activities in our analysis, the analyzed sequences will have more and more longer series of iterated characters, which can be easily compressed and therefore lower the value of entropy per character.

To confirm that, we observe that S follows a scaling law $S(\Delta t_{tp}) \propto \Delta t_{tp}^{\beta}$, where $\beta = 0.75 \pm 0.03$. We can see that this scaling law is only a consequence of the observed distribution of the activity downtimes t_s. This distribution, for activities shorter than 4 hours, has been observed to follow a power law $p(t_s) \propto t_s^{\alpha}$ with an exponent $\alpha \approx -1$. We have generated sequences constituted only by two characters (0 or 1). This choice permit us to reduce the errors of the LZ estimator, that converges slower to the real value of S for larger numbers of characters in the used alphabet. A random character is picked and is repeated r times, with r distributed as t_s. Then another random character is picked and repeated, and so forth. Sequences of these kind, with length equal to the number of minutes in a month, have been generated and shorter sequences, corresponding to the different timeframes Δt_{tp}, have been derived from the original sequence. The entropy of these shorter sequences have been estimated with the LZ algorithm and it follows a scaling low $S_{mc}(\Delta t_{tp}) \propto \Delta t_{tp}^{\beta_{mc}}$, where $\beta_{mc} = 0.73 \pm 0.03$, consistently with the empirical curve.

If we want to avoid these repeated characters, the spatial resolution has to be scaled together with the time resolution, and the complete spatial dynamics of the mobility agents should be analyzed. Unluckily, our data do not permit this fine spatial analysis, and therefore we cannot extract any activity time related feature from our time patterns. However, using a different approach, that is still possible, and it will be presented in the next section.

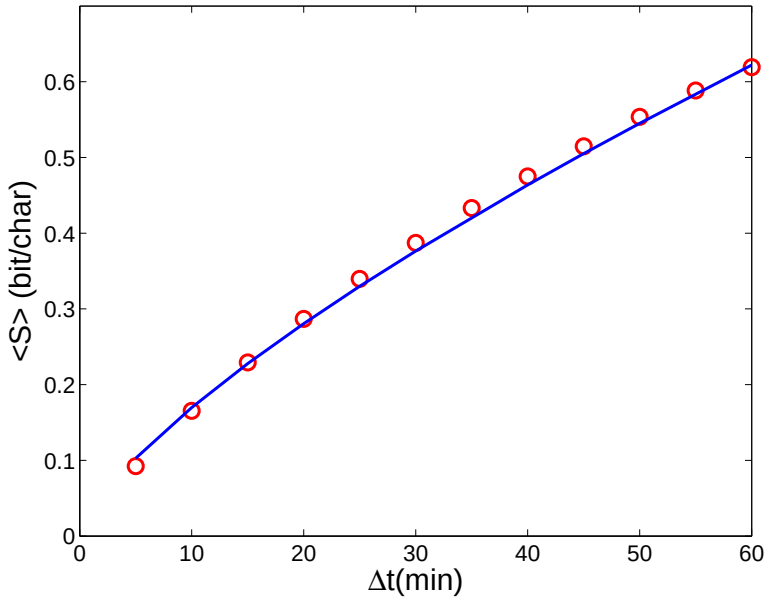


FIGURE 6.4: The scaling of $\langle S \rangle$ can be explained measuring the entropy of Monte Carlo simulated sequences that share the same Benford's Law distribution of the lengths of repeated characters. Thus, the fact that the entropy is small for small timeframes is due only to the progressive growth in length of the sequences of repeated character describing the same activity.

6.4.2 Jump patterns

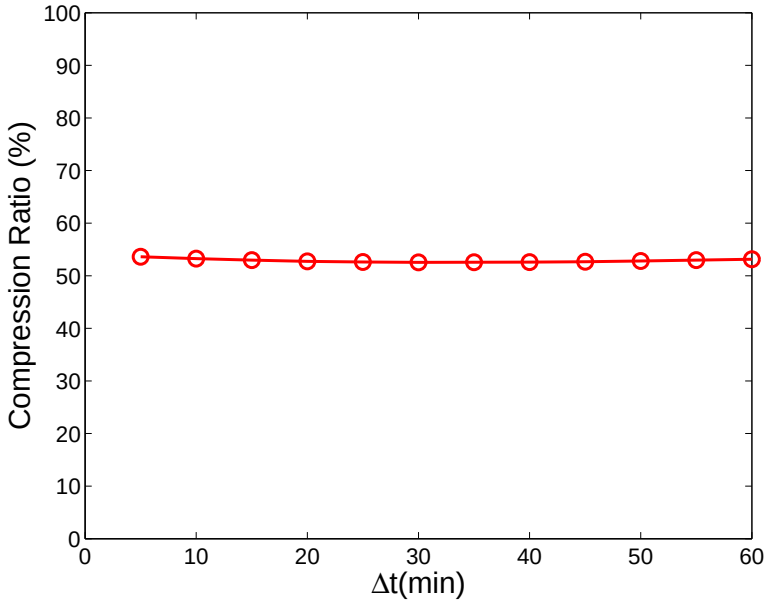


FIGURE 6.5: **(Right picture)** The compression ratio $\frac{\langle S \rangle}{\langle S_r \rangle}$ has no dependency on the threshold value. This suggests a time invariant structure of the correlations including activities shorter than one hour.

We have seen in the previous section that the entropy of time patterns is unsuitable for describing the information carried by human mobility at short time scale. Indeed, in time patterns the fundamental pieces of information describing the movements lie hidden in the transition from one iteration of characters to another. The statistical characteristics of the stops duration play the leading role in determining the value of S. We want to focus our attention to these transitions, and for doing this we analyze the jump patterns, where only the movements are considered.

We use again values of Δt_{jp} ranging from 5 to 60 minutes, that here represent a thresholding value under which stops are excluded from the pattern. Larger thresholds can hardly be considered, as it would produce string too short to be consistently analyzed with the LZ algorithm. Also in this case, the distributions of the three measures of entropy among the different users have their only dependence on Δt_{jp} in their mean value (see fig. 6.2) and we can consider only this quantity in our analysis. In the left graph in fig. 6.5, we may observe a number of differences with the corresponding graph for the time patterns. First, S_r behaves differently, as the way the characters are excluded is different. Second, the values of S_u and S are greater in the jump patterns, as we have neglected all repetitions that were dominating the distribution of character frequencies and are easily compressible. Third, shortening the timeframe and thus introducing new characters, all the three measures of entropy raise. This last observation is straightforward, as introducing new characters we introduce new information. Although, we may notice that the change in S is not great[2], especially as common sense would suggest that the shorter activities, being to be free from organizational constraints, might appear more randomly within the mobility pattern than the longer ones. We would have expected a more steep increase of the entropy caused by the emergence of an asystematic, and thus less predictable, behavior associated to shorter stops. Not only that does not happens, but if we consider the compressibility ratio S/S_r, represented in the right graph in fig. 6.5, we notice that it has only a very weak dependence on on Δt_{jp}. The compression ratio is the measure of the maximum possible compression that can made to the string (it is equivalent to the ratio between the size of a zipped file and the size of the original file): for all the analyzed values of Δt_{jp}, its value lies between 52.6% and 53.6% . The fact that the compression ratio is almost constant clearly indicates that the activities that take 5-10 minutes are as compressible as the one that take 50-60 minutes, i.e. they have the same chance of being part of a repeated sequence that the LZ estimator

[2]Translated in maximum predictability through the inversion of equation 6.2, the values vary from 66% to 71%

uses for the compression. This fact seems to imply that the mobility patterns are somehow invariant with respect of the activity time, at least within the analyzed range (5-60 minutes).

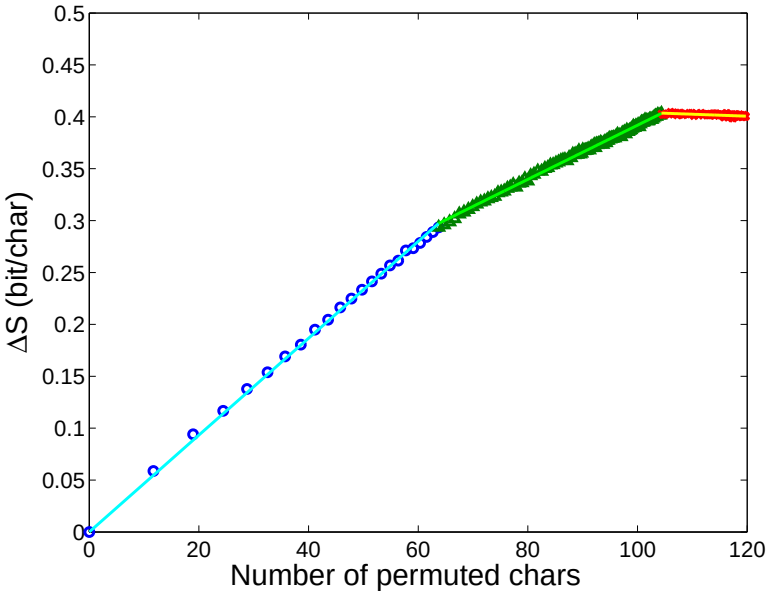


FIGURE 6.6: Performing a progressive shuffling of the characters representing activities which duration is under a threshold $\Delta t'$, we progressively break the correlations in the Jump Patterns. In this figure, we may observe how the average difference in entropy ΔS between the shuffled and un-shuffled patterns grows with the average number of permuted chars with three different slopes. The blue circles represent values of $\Delta t' \leq$115min, green triangles 115min$< \Delta t' \leq$12h and red diamonds 12h$< \Delta t'$.

In order to verify this result, we proceed with a new time-dependent entropic analysis: we compare the values of S measured for jump patterns generated with the minimum threshold value of 5 minutes, after that activities shorter than a given value Δt_s have been randomly permuted. This shuffling procedure breaks the repeated sequences that the LZ algorithm finds and uses for compressing the pattern. The compression due to correlations between consecutive groups of characters is in this way lost and thus the value of S should rise. In the limiting case where all the values are shuffled, the measured value of S is, in principle, equal to S_u. Using this shuffling procedure instead of excluding activities permit us to analyze the activity-time dependency of the correlations without changing the length of the patterns. For this reason, it is possible to extend the range of values Δt_s taken into account. We have here analyzed the values of S with Δt_s spanning from 5 minutes to 24 hours. In fig. 6.6, we show the values of the average difference in entropy $\Delta S(\Delta t') = S(\Delta t_s = \Delta t') - S(\Delta t_s = 0)$ within the sample, plotted against

the average number of permuted characters $N_s(\Delta t_s = \Delta t')$ with the same $\Delta t'$. The greater is the slope of the curve $\Delta S(N_s)$, the faster is the variation in entropy due to the shuffling, and therefore the stronger is the breaking of the correlations due to the shuffling procedure. It is evident how in this curve can be easily divided in three parts, each of which appear to be linear. [3] The best fit with a multiple linear interpolation gives the temporal limits of these parts: 115 minutes (value that, for sake of simplicity in the exposition, we round up to 2 hours) and 12 hours. The slope is greater for the shuffling of the short activities with duration of less than 2h. Then, when we start shuffling the activities longer than 2h with the shorter activities, the variation in entropy falls (the derivative of the 2h-12h part is roughly one half of the derivative of the 0h-2h part). Finally, when even activities longer than 12 hours begin to be included in the permutations, the variation becomes negligible. These three behaviors in the entropy variation permit us to define three classes of activities. The fact that the information introduced with shuffling is simply proportionate to the number n of permuted characters indicates that each class is internally homogeneous. Each character of any value Δt within a given class brings the same increase in entropy when moved from his original position. In particular, this time invariance is valid in the range 5-60 minutes analyzed with the exclusion analysis, which is therefore consistent with this result.

[3] Being this result obtained, for each pattern, with a fixed string length and a fixed number of different characters, we can exclude that this result might be a consequence of some bias due to the LZ entropy estimator.

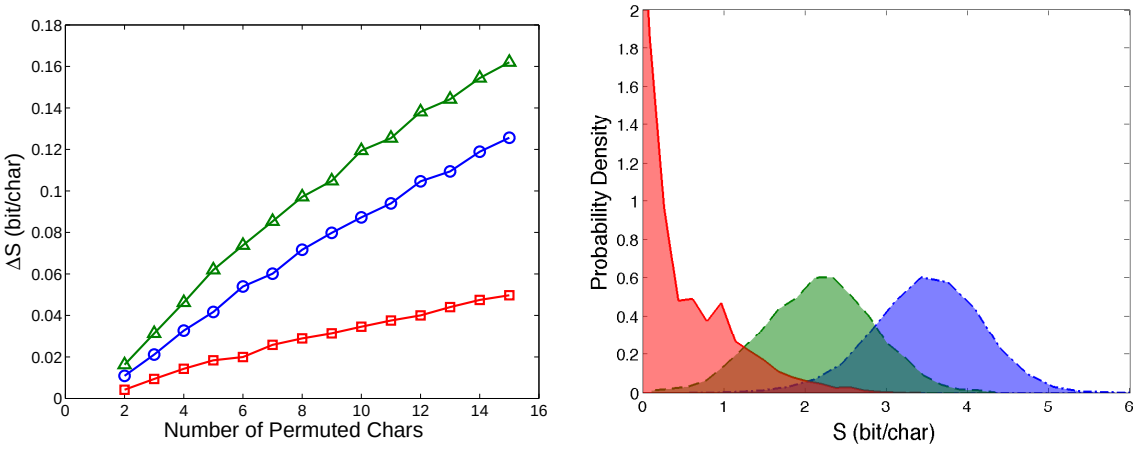


FIGURE 6.7: **(Left picture)** If we shuffle random characters belonging to the same class of activity ($\Delta t \leq$115min: blue circles , 115min$< \Delta t \leq$12h: green triangles, 12h$< \Delta t$: red squares) we point out that the most effective way for breaking the correlations between consecutive stops is shuffling activities of intermediate length. **(Right picture)** The distribution of the Shannon entropies given by the frequencies of the characters in the different classes ($\Delta t \leq$115min: red solid line, 115min$< \Delta t \leq$12h: green dash line, 12h$< \Delta t$: blue dot-dash line). The long activities tend to have a less spread distribution of frequencies (the peak at $S = 0$ for the activities over 12h indicates that frequently only one location (probably home) is visited for such a long time.

Having identified these three classes (that we call respectively short, long and ultra long activities), we can analyze them separately. For doing that, we shuffle a progressive number N_s of characters picked at random within each group, and we measure again the variations in entropy. This analysis has been restricted to the $\approx$ 3800 jump patterns where have been made at least 15 activities for each class. The results, shown in fig. 6.7 left, point out that shuffling characters representing long activities we raise the entropy faster than shuffling the characters of short activities class. The long activities have the smaller derivative, as we could have expected, as only a few locations are visited for such a long time, thus we often shuffle identical characters. This can be observed in fig. 6.7 right, where the distributions of S_u calculated with the frequencies of characters of the three classes are plotted. The curve representing long activities has its values concentrated under 1 bit/char, with a peak at zero, that means that we have usually only one character, representing the overnight stops at home. The distributions of the short and long classes reveal that there is a greater variability in the short activities, while the longer have a less spread distribution of activities, that leads to a smaller value of S_u. This suggests that that differences observed between the derivatives of these two classes in the left picture is even more significant, as the long activities

have a faster rate of growth of entropy with the permutations, even if it is more probable that these permutations are inactive because of the repetition of the same character.

Summing up, if we shuffle first the short activities and then the long activities, the rate $\Delta S/N_s$ relative to the part where long activities are mixed is smaller than the rate relative to the short activities, while if we start shuffling the long activities, the rate is bigger. Besides if we proceed with the shuffling of the ultra long activities after having shuffled the short and the long, the shuffling is ineffective, while the shuffling inside the class have a finite, even if small, effect. We may interpret those facts as hints of a hierarchical structure in the mobility pattern. This hierarchy appears in the repeated sequences that the LZ algorithm recognizes for the compression. Those repeated sequences, representing the individual habits, are more easily broken if we shuffle long stops than if we shuffle the short ones. But, if we shuffle first the short activities, shuffling then the long activities becomes less efficient in breaking these sequences. In our opinion, this can be explained by the assumption that a cluster of activities constitutes a significant part of the repeated sequences where long activity plays a pivoting role between the short activities. Shuffling the long activity breaks more efficiently those sequences, while if we shuffle all the short activities, shuffling then the long activities has not strong consequences. These clusters are embedded, with the possible presence of other correlations, in the middle of an arrangement formed by ultra long activities (representing the cornerstone of the mobility schedule), other hierarchical clusters, other repeated sequences and other activities. We may remark that this interpretation is consistent with our time usage model where the daily schedule is created progressively, starting from the activity with the long duration and progressively using the time left in the timetable (see section 3.1.1).

6.5 Markov processes on network

In this last part of the chapter, we evaluate if the information of the examined jump patterns can be successfully modeled with discrete-time Markov processes on individual mobility networks. In this dissertation, we have already successfully used a Markov assumption in 5.1.2, where we have shown how the exploration of individual networks is related to the topological characteristics of the network resumed by the rank distribution's scaling exponent.

Therefore, we have evaluated, with the Lempel-Ziv estimator, entropies of patterns generated by Markov sources, where the probability matrix is computed from the network adjacency matrix. Two alternative sources have been compared: one where the probability is given by the weights of the undirected mobility networks (see appendix D.1 for the adjacency matrix notation):

$$p^w_{i\to j} = \frac{w^u_{ij}}{\sum_k w^u_{ik}} \tag{6.3}$$

and another where only the topology, intended as the absence or the presence of a connection between the two links, is taken into account:

$$p^t_{i\to j} = \frac{a^u_{ij}}{\sum_k a^u_{ik}}$$

If this second "Topological" process would be able to successfully describe individual mobility, weights would not matter and thus interaction network would carry all the information needed for describing collective movements. As we see in the end of this chapter, this is not the case.

For each network, patterns of 2000 characters have been generated 10 times. This is reducing at minimum levels both statistical error and systematic error due to the entropy estimator. The values of entropy found are represented in figure 6.8 against the entropies measured for the actual patterns.

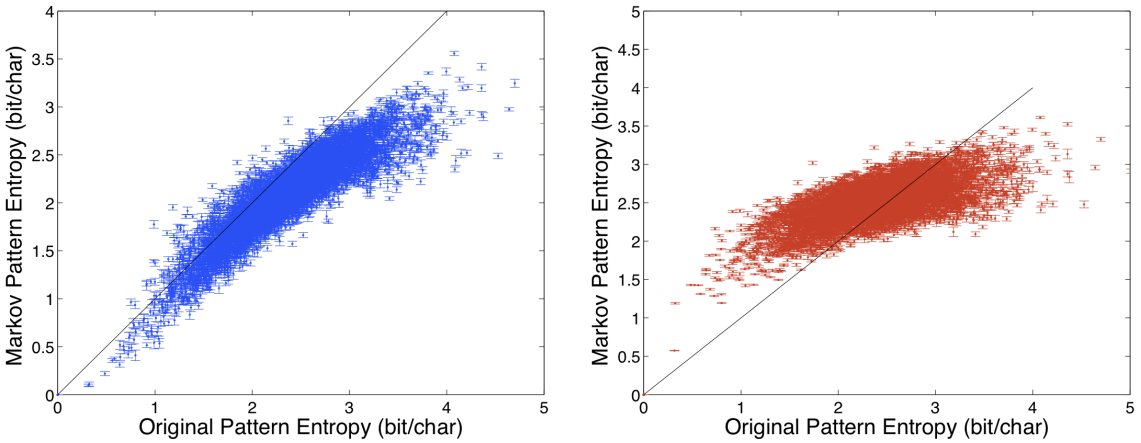


FIGURE 6.8: Relationship between empirical jump entropies and entropies of patterns generated with the $p^w_{i\to j}$ (Left) and $p^t_{i\to j}$(Right). In both pictures the solid black represents the identity.

It appears clear that for low-entropy patterns, Markov processes generated with weights are considerably more similar to real ones. "Topological" entropy seems

to be in some linear relationship far from identity with empirical entropy. On the other hand, highly informative patterns present significant differences in both cases. These difference can be caused by the limited length of the real mobility patterns, that is probably insufficient for the convergence of the Lempel-Ziv estimator to the real value of S if the number of different characters is too high. A comparison between the goodness of "Weighted" and "Topological" entropies in approximating the empirical values is explicit in figure 6.9.

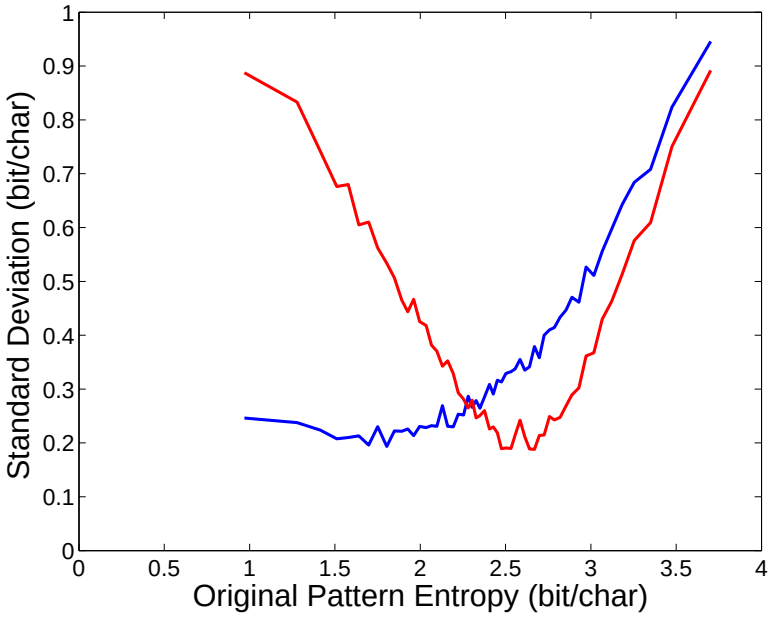


FIGURE 6.9: Standard deviation of the generated pattern entropies from the empirical ones. The blue curve represents deviations of patterns generated with $p^w_{i\to j}$ while the red one represents deviations of patterns generated with $p^t_{i\to j}$. From this graph emerges that the firsts patterns are in good correspondence for a much wider range of values of entropies, while in case of highly informative patterns both processes fail to emulate reality.

Consequently, we can assume that the a Markov process defined by a matrix obtained by the weighted adjacency matrix is reasonably good for explaining the pattern entropy in a wide range of values.

Lastly, we have verified the stability of this consistence introducing a progressive rewiring in the multigraph defined by the weights w^u_{ij} and measuring the consequent increase in the generated pattern entropy. The results exhibit an initial linear trend (see fig. 6.10) with a slope (0.014 bit/char^2) that lies between the empirical slopes measured for activities under 2 hours (0.010 bit/char^2) and activities between 2 and 12 hours (0.015 bit/char^2) relative to the graph of figure 6.7 left[4].

[4]For sake of completeness, the slope for shuffling over 12 hours is 4.3e-03 bit/char^2.

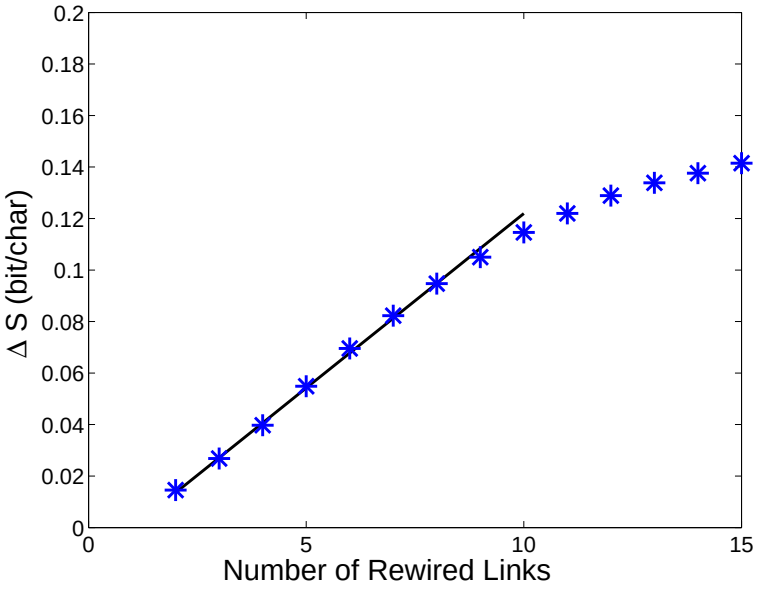


FIGURE 6.10: Effect of progressive rewiring of the individual networks in the values of pattern entropies generated by the transition probabilities (6.3). For a small number of rewired links, the variation grows linearly, as seen for the empirical patterns in fig. 6.7.

This second quantitative consistency grants a greater interchangeability between real and generated patterns. The a-posteriori information carried by the mobility network is therefore can be sufficient for describing the information entropy of the dynamical process that has generated it. Together those topological properties, activity durations are still playing two roles in jump patterns. A major one that we have seen in section 5.1.2, as they limit in the size of the pattern's alphabet. Second, subtler, in the hierarchical structures that are suggested by the time-shuffling study of last paragraph.

Chapter 7

Mobility on the street network

As we already discussed in chapter 4, human mobility is constrained by energetic consumption due to the effort of traveling[37]. Once the transportation mean is chosen, time, rather than distance, appears to be the quantity used to evaluate remoteness or proximity. Travel-time defines a metric that depends on many factors, notably driver characteristics and traffic conditions. Driver's decision dynamics can be probabilistically modeled using the concept of utility function[55]. As the benefits of a journey decreases with the travel time, utility is at first order proportional of with its inverse function $\bar{v}/l$[56]: with a given distance l to cover, utility is therefore linear with the average speed $\bar{v}$. This chapter intends to analyze the features of the travel-time metric, which is continuously deformed by traffic conditions, and how those features influences drivers behavior.

7.1 Space-Time Relationship

For comparing the fixed spatial metric defined by the street network with the temporal metric, we study the average speed of the vehicles' trajectories in the Florence dataset. It has been observed[57] that the relationship $l = \bar{v}t$ is valid only for long trips, while it is not consistent when the distance covered l is short. In fact, the short paths' vehicular micro-dynamics is dominated by the microscopical interactions with the street network, and these interactions cause a continuous alternation of accelerations and decelerations. Only for the longest trips these fluctuations become negligible and an average speed independent by l or t is well defined.

To point out the regime dominated by microscopical interaction, we consider only the trajectories belonging to a daily cycle, i.e. a daily mobility patterns starting and ending in the same location. These cycles represents a typical day of mobility with night rest at home. This choice automatically excludes "diffusive" movements and focuses more on urban mobility. The microscopical interactions can be assimilated in a stochastic model[57] where the succession of accelerations and decelerations are described by white noise:

$$\begin{aligned} \dot{v} &= \sigma\xi(t) \\ \dot{l} &= |v| \end{aligned}$$

with $\langle\xi\rangle = 0$ and $\langle\xi(t) - \xi(s)\rangle = \delta(t-s)$.

It is common knowledge that v follows a Wiener:

$$p(v,t) = \frac{1}{\sqrt{2\pi\sigma^2 t}} \exp\left(-\frac{v^2}{2\sigma^2 t}\right)$$

Therefore, the average speed is:

$$\langle\dot{l}\rangle(t) = \langle|v|\rangle \simeq \sigma\sqrt{t}$$

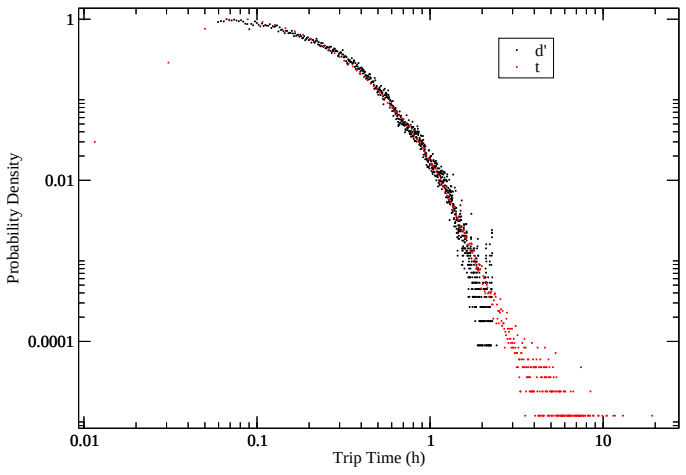


FIGURE 7.1: Comparison between the travel-time distributions and the distribution of the quantity d' obtained from transformation (7.1). Data are from the Florence dataset with restriction to daily cycles.

It is important to highlight that this relationship between average speed and travel time can be used, like in figure 7.1, to convert distances in temporal measures with the transformation:

$$d' = \frac{d^{\frac{2}{3}}}{K} \tag{7.1}$$

With $K = \sigma^{\frac{2}{3}} \approx 9.3\frac{Km^{\frac{2}{3}}}{h}$, we obtain a good agreement between the two metrics. We remark that the good correspondence shown in figure 7.1 is obtained only if not-cyclic days are excluded from the analysis.

To better understand the space-time relation, we have studied the variance of the average speed as a function of the trip length. In the figure 7.2 we plot the result for the whole Emilia-Romagna dataset: the data show a power law increase of the variance for very short trips and a relaxation to a stationary condition for trip lengths greater than 8 km: the stationary variance corresponds to a *rms* $\sigma_0 \simeq 10$ km/h in the speed distribution (the red line show an interpolation of the experimental data).

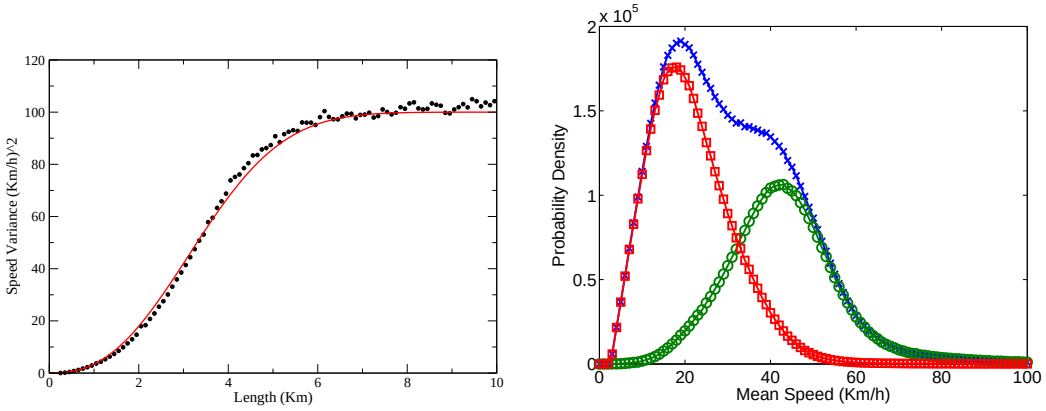


FIGURE 7.2: **(Left picture)** Average speed variance as a function of the trip length: we have computed the average speed for a given trip length using the GPS data of the Emilia-Romagna dataset with a discretization step of 100 m. The continuous line is a data interpolation using the function $\sigma^2(l) = \sigma_0^2(1 - \exp -(l/b)^{5/2})$ with $\sigma_0 = 10$ km and $b = 3.8$ km. **(Right picture)** Average speed distribution for the recorded trips (blue curve): the distribution can be decomposed into the sum of two distributions considering the trips whose length is ≤ 5 km (red curve) and the remaining ones (green curve).

However, considering the average speed distribution for all trips, it is possible to point out the two different typologies of trips: the short trips $l \leq 5$ km with an average velocity $\simeq 20$ km/h and a small variance and the longer trips with a distribution centered at $\simeq 45$ km/h and with a rms $\sigma_0 \simeq 10$ km/h (fig. 7.2 left).

We remark that the two trip typologies are not directly related to the exponential and power law behavior in the trip length distribution (see chapter 4) since the power law behavior can be detected considering trip longer than 20 km. Instead, the length-time relationship (7.1) can be the cause of the apparent reduced short trips' suppression effect shown in the Total Mobility distribution with respect to Travel-Time Budget (confront the shape of the curves on the two sides of fig. 4.5). In fact, even in mobility with short travel-time is suppressed, mobility with short travel distance is still possible because it tends to be slower and therefore takes a relatively longer time.

7.2 The Bike alternative

The slower speed of short trips represents a disutility that, together with the time spent directly and indirectly preparing from locomotion, represents one of the hidden costs of transportation[58][1]. Moreover, traveling at low speed because of the continuous accelerations and decelerations is a source of stress, and stress rises the energy consumption influencing the Travel-Time Budget. For those reasons, it can become the optimal solution switching for short tripsbetween cars and other forms of transportation.

To quantify this, we take as example the city center of Milan, a city where the stress from driving is notably high[59]. We have restricted the analyzed area to the region within the "Circonvallazione esterna", a ring of avenues with diameter of $\approx$ 6Km. We have selected only the trajectories with both starting and stopping point within the area, a portion representing the 47% of the total trajectories involving the area. The distribution of the average speed in this selection is plotted in figure 7.3.

[1]Disclosed costs are, for instance, fuel and parking space, the buying price of the vehicle, the expenses for its maintenance or the costs for the upkeep of the street network that are paid by the community.

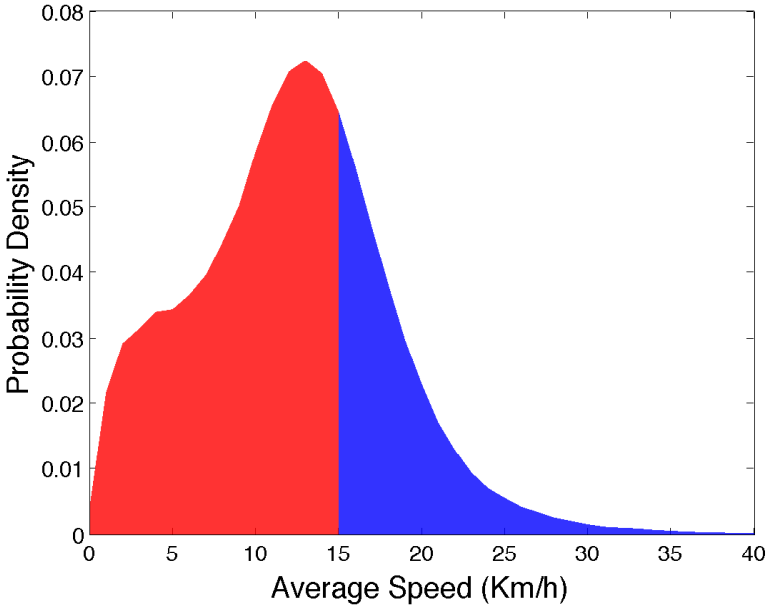


FIGURE 7.3: Distribution of $\bar{v}$ for trips within the "Circonvallazione esterna" in Milan. In red is highlighted the share of trajectories slower than 15 Km/h, representing the 73% of the total.

We observe that over the 70% of the trips inside the "Circonvallazione esterna" are slower than 15 Km/h, which we pick as reference measure because it is a standard speed for bikes. That means that the larger part of the car movements in the central area of Milan could would be faster if done by bike, even without considering the time spent from the parking place to the actual destination (time that tends to be greater for cars' parking). If we take into account also walking times from the activity to the car, cars becomes even less attractive for very short distances[42]. Also bike movements are probably limited by crossroads, traffic lights and the lack of infrastructure in a way similar to car mobility, and therefore a relationship similar to (7.1) can be introduced for this transportation mean: also in this case the average speed is not constant and for longer distances fatigue has to be taken into account as well. But, the fact that cars are chosen instead of bikes reflects that other factors, like comfort or safety, are taken into consideration for modal choice.

This quantitative analysis confirms that bikes are a perfect alternative to cars for urban mobility. If the dis-utilities given by insufficient cycle facilities are reduced with specific developments policies, the speed advantage can prevail, triggering a mode shift with benefit for both bikers and drivers.

7.3 Route Assignment

7.3.1 The *oeconomicus* driver

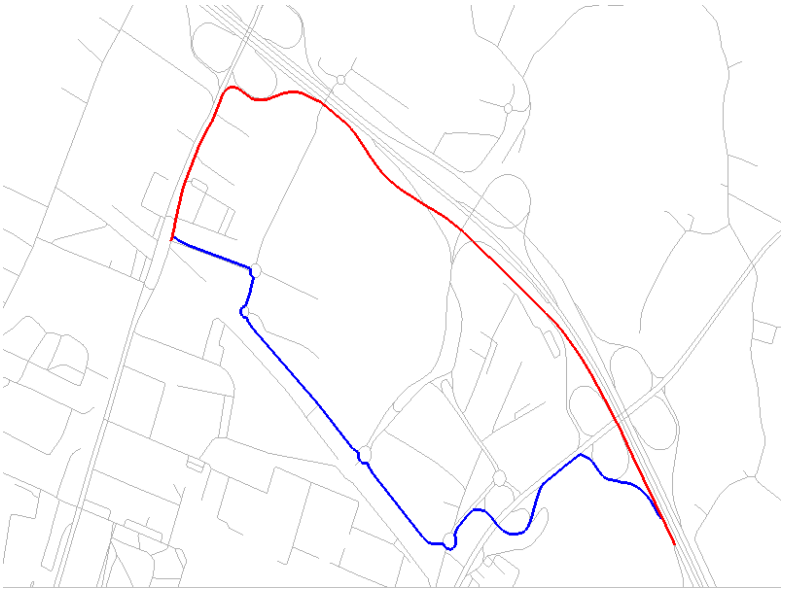


FIGURE 7.4: Two alternative routes in Bologna: the northern option is longer but takes advantage of a ring road while having therefore higher average speed; the southern option is shorter but involves numerous roundabouts.

In transportation theory, a driver as considered similar to the *homo oeconomicus*. Choices among alternatives, like what mode of transportation select or what is the best trajectory from A to B, can be explained with the multinomial logit model[55]. In particular, for the case of route choice, travel time can be used as cost function[56], so that a driver moving from x to z at a traffic node opts for the alternative route $k \in \mathcal{R}$ with the probability:

$$P_k(t) = \frac{\exp(-T_k(t)/T_0')}{\sum_{k' \in \mathcal{R}} \exp(-T_k(t)/T_0')} \tag{7.2}$$

where $T_k(t)$ is the travel time expected at time t for the route k and T_0 a suitable proportionality constant between travel time and utility:

$$U_k(t) = \frac{T_0}{T_k(t)} \tag{7.3}$$

As traffic situation and expected travel times may significantly change while the next node k is not yet reached, may be in some cases useful to re-estimate the

travel time at each node

$$T_k(t) = T_t(x,k) + T_t(k,z) \tag{7.4}$$

More generally, the multinomial logit model permits to estimate the probability $P(x|s,B)$ of choosing the option x from an alternative set B given the measured attributes vector s (where both individual parameters and external information lie). The model is based on several assumptions[60]. Some of the most relevant are:

Independence of Irrelevant Alternatives

$$\frac{P(x|s\{x,y\})}{P(y|s\{x,y\})} = \frac{P(x|s,B)}{P(y|s,B)} \tag{7.5}$$

;

Irrelevance of Alternative Set Effect

$$U(s,x,z) = u(s,x) - u(s,z) \tag{7.6}$$

taking z to be a "benchmark" member of B;

Linearity of $u(s,x)$

$$u(s,x) = \sum_{k=1}^{K} \theta_k u^k(s,x) \tag{7.7}$$

where $u^k(s,x)$ are specified numerical functions and the θ_k unknown parameters.

Under these and other assumptions the multinomial logit model probability:

$$P(x|s,B) = \frac{\exp(U(s,x,z))}{\sum_{y\in B}\exp(U(s,y,z))} \tag{7.8}$$

is consistent with a model of individual behavior where the actual utility function is the sum of two effects:

$$u^* = u(s,x) + \epsilon(s,x) \tag{7.9}$$

where u is non-stochastic and reflects the "representative" tastes of the population, and ϵ is stochastic and reflects the idiosyncrasies of a particular individual in tastes for the alternative with attributes x and the randomness of a particular choice made by the individual.

7.3.2 Wardrop's Equilibrium

From the rationality and total knowledge assumed for drivers' decisions, together with the fact that when a lot of vehicles try to use the same road, the road becomes congested and travel-time increases, come as consequences the Wardrop's principles[61].

The Wardrop's first principle states that a traffic flow over a street network is at equilibrium (a Wardrop equilibrium) if no user can reduce his cost by switching from his current path to another one connecting the same origin-destination pair. This flow equilibrium represents a steady state in a system where the flow is generated by a very large number of infinitesimal users. It has been demonstrated that a Wardrop equilibrium is equivalent to a Nash equilibrium for congestion games with a large number of players[62]. In this kind of game, it is assumed that each driver is acting in a purely selfish manner, and will use the minimum-latency path from its source to its destination, given the congestion caused by the rest of network users. These congestion games are also potential games, because a potential function can always be defined[63].

The Wardrop's second principle states that at equilibrium all flow paths between a given source and destination have equal, and smallest possible latency. However, this traffic assignment could be far from the optimal assignment minimizing total latency[64]. In this game theoretical framework, it is even possible that improvements in traffic facilities have negative effects due to lack of coordination of drivers, an effect called Braess's Paradox[65].

Summing up, the Wardrop's Principles are based on three strong assumptions:

- system being in a steady state;
- rational drivers are aiming to travel time minimization;
- all players have complete information of the system state.

It is our interests to check if a real urban traffic system can be actually observed at Wardrop equilibrium. In the negative case this can be a consequence, for instance, of the incomplete information about the system state or a less-rational behavior like inertia in switching from of the usual route even if it has an higher cost[2].

[2]This study has been conducted under the supervision of Prof. Dirk Helbing.

7.3.3 Empirical falsification of Wardrop's principles

In order to validate the Wardrop's principles, it is necessary to confront the trajectories followed by different drivers at the same time. Therefore, we have to bin our data in timeframes, each identified by its center in t and of width ΔT. Then, we may identify a particular trial $n = [o, d, t]$ for each origin-destination couple $[o, d]$ and each timeframe t. In a particular trial there will be J_n different routes labeled with k chosen by M_n drivers. Each driver m have incurred a travel cost τ_{kmn}, that we may associate with the travel time of route k in trial n. The analysis has been done considering the actual travel time τ_{kmn}. For the 2nd principle, in the same trial, for all k and m the values of τ_{kmn} should be identical.

Two samples of the Italy dataset with the mobility in the city of Bologna and Florence have been chosen. These municipalities are of similar size (respectively 380.000 and 370.000 inhabitants), but due to a different commercial diffusion of the GPS device, the two samples are of different sizes: 1.35 million trips in Bologna, 0.79 million trips in Florence. Even with those large sizes, only a limited number of trials have been found with information sufficient to make a statistical comparison between travel times on two alternative routes. The important bottleneck comes from the low recording frequency of the devices: with only one datum every 2 Km all the movements between one point and another have to be guessed with path reconstruction. To have complete certainty that two trajectories are actually different, it is necessary to consider only trips with at least 3 matched points, which requires a minimal length > 2 Km. Moreover, picking timeframes 15 minutes, we have to find at least 3 cars that in that time span have chosen at least two different roads between the same couple of links in the street network. The need of 3 cars is for computing a confidence interval for the travel-time in at least one of the routes in the trial, which is necessary for evaluating the consistency of the travel-time in the other route. In almost all cases, only one route is commonly used while the alternative choice is relatively rare. In fact, the flow of a car with a device installed 15 minutes is approximately equivalent to 200 car/h along the entire alternative route. This is involuntarily focusing our analysis on rush hours, when these flows are possible.

is the two data samples, we have identified 301 timeframes with enough statistics over 31 different route pairs. Each trip has been used only once in the analysis, so when it belonged to more than a route pair, the one covering the longest has been chosen. Our finding is that Wardrop's equilibrium is significantly violated 50.5% of times. The result is similar for the two samples (53% in Bologna and 46%

Florence and is consistent in all hours of the day. In figure 7.5 left, we represent the percentage of Wardrop's equilibrium violations as a function of the hour of the considered timeframe: even if statistical errors do not make the differences significant, the graph is suggesting that at rush hours the equilibrium is violated more than usual.

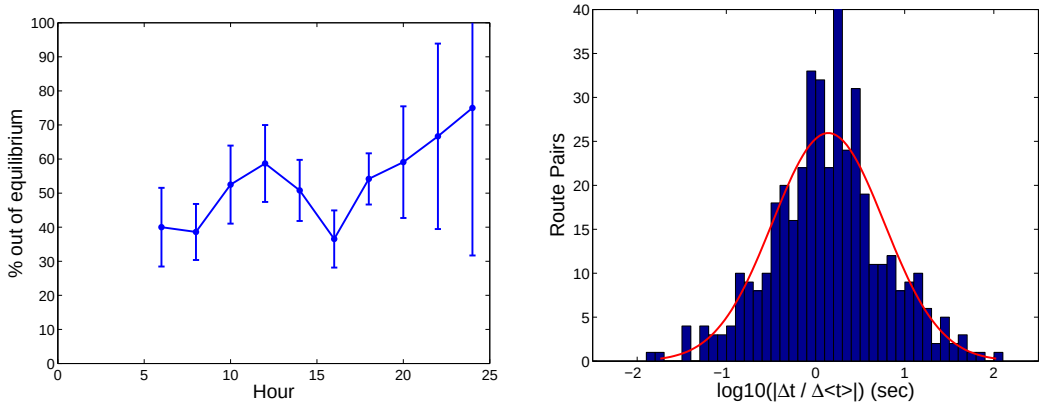


FIGURE 7.5: **(Left figure)** Hour dependency of the percentage of Wardrop's equilibrium violations. **(Right figure)** Distribution of the relative fluctuation around the average of travel-times in the selected trials: the red line suggests a lognormal interpolation.

A possible explication for the deviation from equilibrium would be that the cost function involves other alternative to travel time. Possible path costs can be function of[66]:

- length;
- angular change (or, alternatively total square curvature of the trajectory);
- number of turns (or eventually only number of turns left);
- travel-time variance (greater variance can be interpreted as higher risk).

To compensate differences in travel-time, an additional term should be anti-correlated with them. However, no significant anti-correlation has been found with functions of the costs present in the list[3].

A second analysis over the average speed of the routes has been done. Those values may better represent the value for the travel cost as the drivers can anticipate it.

[3]This has been verified supposing either a linear or a logarithmic travel-time cost. A logarithmic cost can be in principle be consequence of the Weber-Fechner's law (section 3.3).

Indeed, travel-times τ_{kmn} associated to the effective realization of the trip from o to d by the driver i are precise and reliable, but in principle they cannot be known by the drivers when the route is effectively assigned. For average travel-times, we find that only 2 of the 31 route couples have average times significantly different. Therefore, average travel time can be suggested as a consistent cost function. This implicate that most of the drivers can only rely on average knowledge of the travel times.

Finally, we want to point out that, for the same route, all observed travel-times are log-normally distributed (see fig. 7.5). This cannot be a consequence of experimental errors, which are for our measures Gaussian. It is possible to suggest that also this distribution can be related to a Weber-Fechner's law of perceived times. This non-linear perception could be fundamental for decisions, as Weber-Fechner's law has been identified as a possible unique assumption needed for obtaining in a multinomial logit function in quantitative decisions[55].

Chapter 8

Methods for Traffic Data Assimilation

Thanks to the commercial diffusion of GPS technology, we have now access in real time to information on traffic condition, which are not limited anymore to particular places where measure instruments and cameras are installed, but are available over the entire street network. The information, coming form navigators, cell-phones or black boxes installed on vehicles, is extremely precise but has two important limits. First, it is associated to the mobility of a single vehicle, which might not be a good representative of all the traffic flow. Second, it can be too dispersed over time and space, so that in some location no recent knowledge on traffic condition are at disposal even for long periods of time. In this chapter we present the data assimilation methods we have applied to our path reconstructions (see section 2.3) in the real-time traffic data assimilation software developed for the industrial project Pegasus - Industria 2015[67].

8.1 Kalman Filter

We are have information about the average speeds at various edges of the street networks, obtained by the reconstruction of trajectories followed by single drivers[12]. Data at our disposal are not homogeneously available either over time, as they are abundant in rush hours and rare during the nights, or over space, as we can have from the 100 data/hour of main roads to the total absence of information over one month of recording. For instance, on the whole Italian street network, only the

3.6% of the total length have a monthly flow of cars equipped with Octo Telematics GPS device greater than 3000, corresponding to ≈ 4.5data/hour and thus to a real flow of ≈ 150 vehicles/hour[68].

Our objective is to give an optimal estimate to the values of speed and flow at a given time in those roads where we have a reasonable quantity of data. As our practical goal is to develop software able to work in real time, we cannot use the *a posteriori* knowledge of the future state of traffic, but we can only rely on the last updates, the previous traffic conditions and historical information. Moreover, a further factor to be considered is computation time, as data elaboration for real time purposes should be constraint to an update pace from 5 to 15 minutes.

The solution we propose here is based on Kalman Filter. Kalman Filter has been chosen as it permits to:

- enrich the information on a given time with the state of the system in the preceding times;
- give a reasonable estimate even in absence of data;
- reduce the errors due to the variability of data;
- improve the estimate for a given road with information received in the neighboring area.

In particular, we have implemented the Discrete Linear Kalman Filter[1] with the additional assumptions:

- $\mathbf{Q}_k = \sigma_f^2 \mathbf{I}$
- $\mathbf{R}_k = \sigma_o^2 \mathbf{I}$
- $\mathbf{H} = \mathbf{I}$

which means that both forecast and observational errors are identical and independent for all analyzed roads and, at the same time, the observed values maps perfectly the fields considered for the model state.

[1] An explanation of this method is found in appendix F.

Kalman equations become (vector notation is now not necessary as the problem has become diagonal):

$$\begin{aligned}
\text{Forecast Step}: \quad x^f_{k+1} &= \phi_k x^a_k \\
P^f_{k+1} &= \phi_k P^a_k \phi^T_k + \sigma^2_f I \\
\text{Analysis Step}: \quad x^a_k &= x^f_k + K_k(z_k - x^f_k) \\
P^a_k &= (I - K_k)P^f_k \\
\text{Kalman gain}: \quad K_k &= P^f_k(P^f_k + \sigma^2_o I)^{-1}
\end{aligned}$$

Dividing the two equations about variances P by σ^2_f we obtain the equation for $\frac{P}{\sigma^2_f}$:

$$\begin{aligned}
\tfrac{P^f_{k+1}}{\sigma^2_f} &= \phi_k \frac{P^a_k}{\sigma^2_f}\phi^T_k + I \\
\tfrac{P^a_k}{\sigma^2_f} &= (I - K_k)\frac{P^f_k}{\sigma^2_f} \\
K_k &= \frac{P^f_k}{\sigma^2_f}\left(\frac{P^f_k}{\sigma^2_f} + \frac{\sigma^2_o}{\sigma^2_f} I\right)^{-1}
\end{aligned}$$

Redefining $\tilde{P} = \frac{P}{\sigma^2_f}$, Kalman equation are:

$$\begin{aligned}
\text{Forecast Step}: \quad x^f_{k+1} &= \phi_k x^a_k \\
\tilde{P}^f_{k+1} &= \phi_k \tilde{P}^a_k \phi^T_k + I \\
\text{Analysis Step}: \quad x^a_k &= x^f_k + K_k(z_k - x^f_k) \\
\tilde{P}^a_k &= (I - K_k)\tilde{P}^f_k \\
\text{Kalman gain}: \quad K_k &= \tilde{P}^f_k\left(\tilde{P}^f_k + \frac{\sigma^2_o}{\sigma^2_f} I\right)^{-1}
\end{aligned}$$

where the ratio $\frac{\sigma^2_o}{\sigma^2_f}$ is a free parameter, which has to be experimentally estimated. A deeper analysis of the asymptotic properties of this 1-Dimensional Discrete Linear Kalman Filter can be found in appendix G.

The assumptions leading to this version of the Kalman Filter are not hold if the model is not diagonal, i.e. it uses, to estimate the future state of a given arc, information of the states of other street arcs. For this more elaborated models (such as the one described in section 8.3) is computationally lighter to use the 3DVar method (see appendix H), where the inversion of the $n \times n$ matrix (with

n the number of analyzed street arcs) necessary to compute the Kalman gain is replaced by the optimization of a n-dimensional function.

8.2 Evaluation and calibration of two simple models

There are many possible forecast models developed in traffic theory[3]. Fluid-dynamics models have been implemented for velocity data assimilation on highways[69]. Although, we doubt that a similar approach can be successfully applied at urban scale. On the other hand, models like the gravitation or the radiation model[36], can only describe mean fields and not transitory phenomena.

Therefore, we have chosen to start implementing two extremely elementary models, whose role is only to propagate information of the past in the current state analysis. The first model is temporal persistence ($\phi_k = 1 \quad \forall k$). The second has two regimes: in absence of recent data on a particular street arc it converge exponentially with a relaxation time τ to a value characteristic of the street, while persistence is used again if new data have been received. The characteristic value can be either the empirical speed of that street when it is free from traffic, the speed limit or the average speed.

We have information from a different data source on cars' speed and flow in the highway denoted as "Strada a Grande Percorrenza Firenze-Pisa-Livorno (FI-PI-LI)" in the time span covered by the Florence dataset. These data have been taken by coils immersed in the street in different points along the highway. One particular section (close to the exit from a freeway) presented marked speed variability and therefore 8 days of speed and flow records relative to this location have been chosen for the calibration of the $\frac{\sigma_o^2}{\sigma_f^2}$ ratio. Indeed, these data, averaged in 5 minutes timeframes, can be reasonably assumed as correct.

The calibration has been performed confronting coil data with the Kalman analysis results obtained from path reconstruction data. As initial conditions for the Kalman analysis we took the average value of the first 4 hour of GPS records ($78\frac{Km}{h}$.). To this value it has been arbitrarily associated a variance equal to the square of the mid-range of the values of speed for the same time span ($237\frac{Km^2}{h^2}$). For the exponential convergence model, besides the ratio $\frac{\sigma_o^2}{\sigma_f^2}$ also the relaxation time τ has been subject to calibration. Moreover, we have defined two different metrics for the measure of the quadratic deviations between coils' "real" data

and GPS-Kalman reconstructions. In one case, all timeframes have been equally weighted as long as there have been measures of speed in the coils' dataset. In another, deviations have been weighted with the car flow measured by coils in that timeframe. In the following, we call simple the first metric and flux metric the second one. The flux metric has the advantage of focusing on rush hours and moments of congestion, and is therefore more interesting for the practical use of the filter, while the simple metric is a better measure of the overall Kalman reconstruction consistency.

8.2.1 Persistence

As we see in figure 8.1, for the persistence model the optimal average error we obtain is less than $10\frac{Km}{h}$ for both metrics. The optimal value for $R = \frac{\sigma_o^2}{\sigma_f^2}$ is 90 with simple metric and 80 with flux metric.

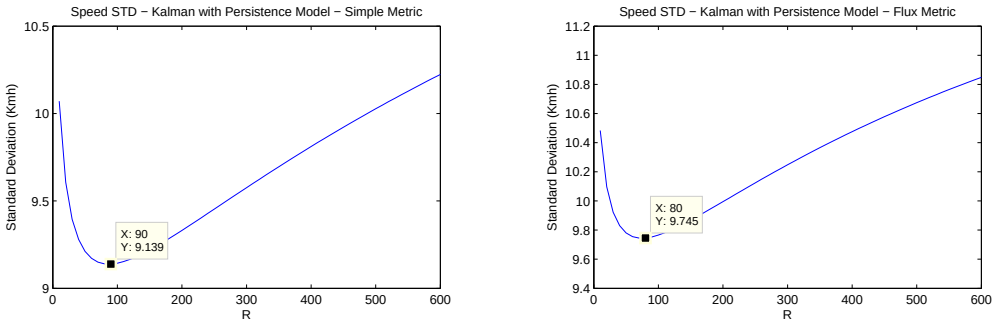


FIGURE 8.1: Relationship between average estimation error and the parameter $R = \frac{\sigma_o^2}{\sigma_f^2}$. The left picture shows errors evaluated with simple metric, while in the right picture errors are evaluated with flux metric. Note that in both cases with values of R ranging many orders of magnitude, errors have variation within the 10%.

Furthermore, we notice that in the range of the estimated optimal parameters, errors present a weak dependency by deviation from optimum: reasonable errors are kept for different order of magnitude of the parameters. So, even if we have calibrated on only one road, these values would probably give acceptable results also in very different conditions.

The result of the analysis with $R = 90$ is shown in figure 8.2.

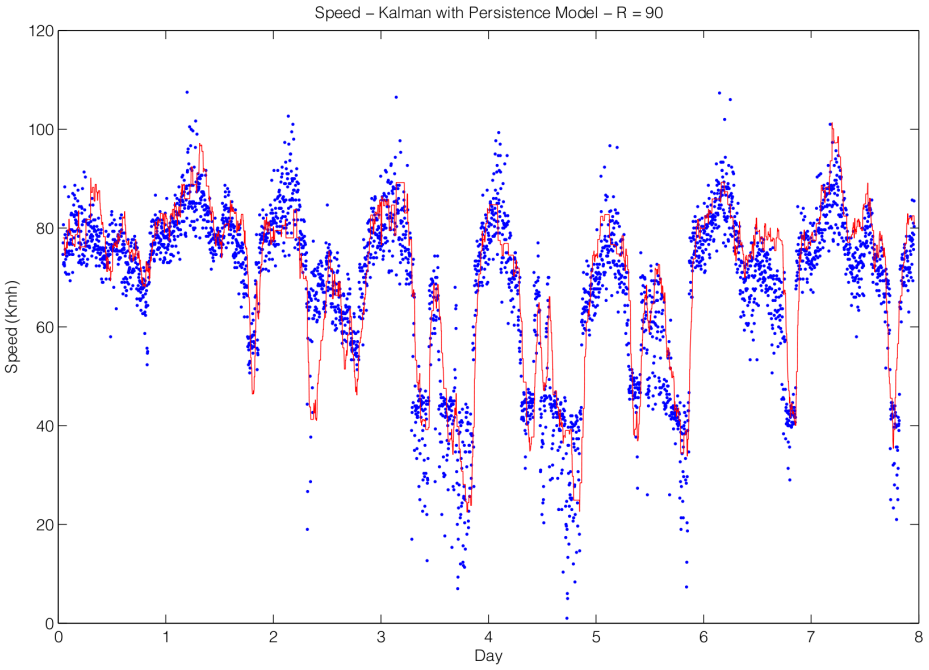


FIGURE 8.2: Comparison of the Kalman analysis with persistence model on path reconstruction speeds (red line) with coils' data (blue points).

The Kalman analysis is characterized by a latency of ≈ 20 minutes with respect to coils data. The numerical value of the latency has been evaluated maximizing the correlation between the two signals with shifted times. This effect is typical of the Kalman filter, as more than one datum suggesting a departure from the expected values has to be fed to the algorithm for considering that deviation trustworthy and not a random fluctuation. Nevertheless, the use of a persistence model is unquestionably increasing this effect. In figure 8.3 we zoom on a smaller time interval to show more clearly this latency.

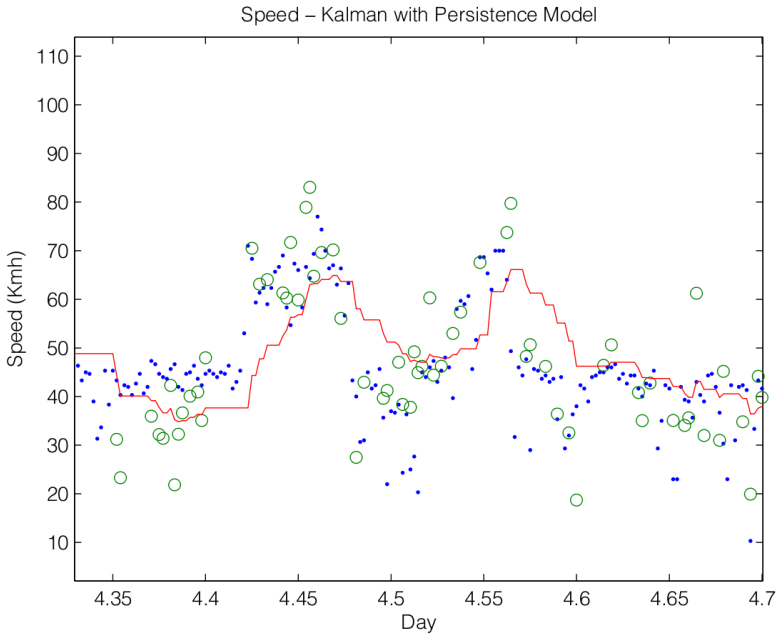


FIGURE 8.3: The Kalman analysis with persistence model presents a latency of $\approx$ 20 minutes with respect to both coils data (blue points) and original GPS speeds data (green circles).

8.2.2 Exponential convergence

For the calibration of the exponential convergence model we have first chosen the average value of speed measured in all 30 days as characteristic converge value, because it gives better statistical consistency with coil data than the speed limit or free speed, particularly during rush hours. For the analyzed section and timespan, this speed was 61 Km/h. We find that with the simple metric we obtain minimal deviations for $\tau = 6h$ and $R = 70$, while for weighted metric for $\tau = 2,5h$ and $R = 60$ (figure 8.4).

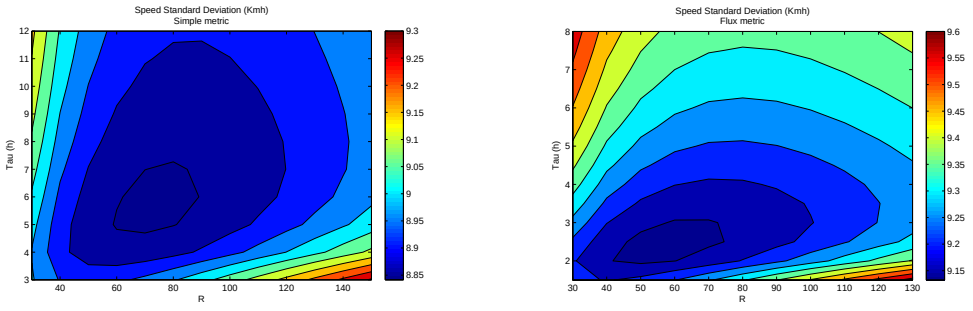


FIGURE 8.4: Relationship between average estimation error and the parameters $R = \frac{\sigma_o^2}{\sigma_f^2}$ and τ. The left picture shows errors evaluated with simple metric, while in the right picture errors are evaluated with flux metric.

Also in this case the average error is far below 10 Km/h and has not a strong dependence with the chosen parameters. With the optimal values for the simple metric, $\tau = 6h$ and $R = 70$, we obtain the Kalman analysis of figure 8.5:

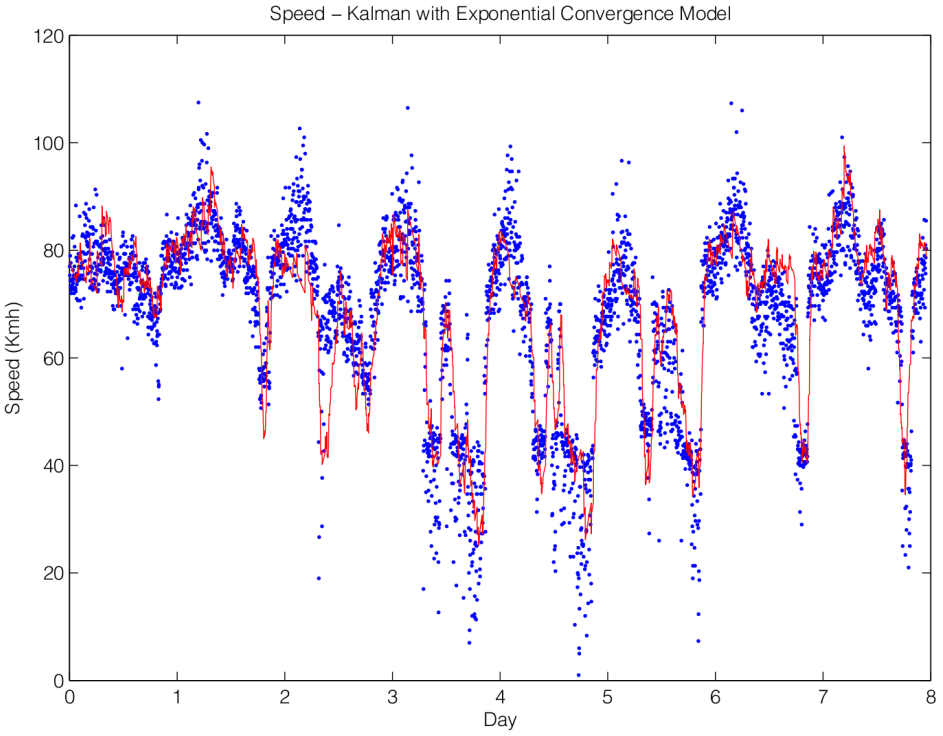


FIGURE 8.5: Comparison of the Kalman analysis with exponential convergence model on path reconstruction speeds (red line) with coils data (blue points).

The exponential convergence model reduces the measured latency to 8 minutes from the 20 of the persistence. But, above all, its principal advantage is to make the analysis converge to an historical value in case of total absence of data update. At the same time, it is also reducing the average error. Furthermore, we have verified that even with significantly less data inputs the value of the error does not have remarkable changes: this is probably a consequence of the weak dependency of the resulting error from estimated parameters.

8.2.3 Flow Analysis

Together with speed, cars flow is a fundamental quantity in traffic analysis. For instance, road network performance can be measured by its ability to carry a certain amount of traffic kinetic energy, namely the product of flow and its speed[40].

To measure real flows from our sample of cars with the GPS device installed, we first compute the average ratio between our flow data and real flows measured with coils. The value found for the analyzed section is 115, corresponding to a device commercial diffusion at the time of 0.87%.

Our flow data are extremely poor: the number of cars passing in a 5 minutes time-frames lies in the range 0-5. Those values are multiplied by 115 to be confronted with coils' data. We have used the persistence model to filter this signal, finding an optimal value of $R = 1800$ with an quadratic deviation of 330 car/h. We have that R is greater for flows than for speeds because data are bringing only a small amount of information. Errors dependency with R and the optimal analysis are displayed in figure 8.6.

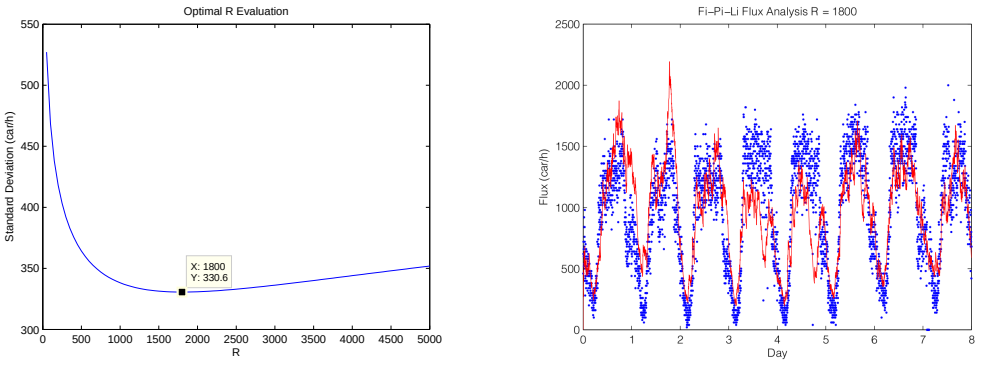


FIGURE 8.6: **(Left picture)** Relationship between average flow estimation error and the parameter $R = \frac{\sigma_o^2}{\sigma_f^2}$ with errors evaluated with simple metric. **(Right picture) Comparison of the Kalman analysis with exponential convergence model on path reconstruction flows (red line) with coils data (blue points).**

8.3 Statistical Traffic Forecast: Application to the Grande Raccordo Anulare

The two forecast models implemented so far in this chapter are not doing, in reality, any real forecast. In this section we try to apply a statistical forecast method developed in meteorology[70]. This method, which detailed description is given in appendix I, defines a linear forecast model minimizing statistical errors given a set of somehow correlated time series. Noise amplification is reduces taking advantage of a projection over Empirical Orthogonal Functions (i.e. a Principal Component Analysis).

In particular, we have tested the method over a set of 410 time series representing the mean speeds, evaluated in 15 minutes timeframes, along the Grande Raccordo Anulare (GRA). The GRA is a ring shaped orbital motorway that encircles Rome. In particular we have isolated the external track and we have placed the point $s = 0$ after exit 25 (Laurentina) and then taken the first half of the month of May 2010 for the calibration of the prediction formula and the second part of the month as a control sample.

8.3.1 Principal Components Analysis

Over this set of signals we have calculated the time covariances and diagonalized the covariance matrix. As can be seen in figure 8.7, the first six functions carry the 95% of the variance, while the following bring less than 1% of the variance each.

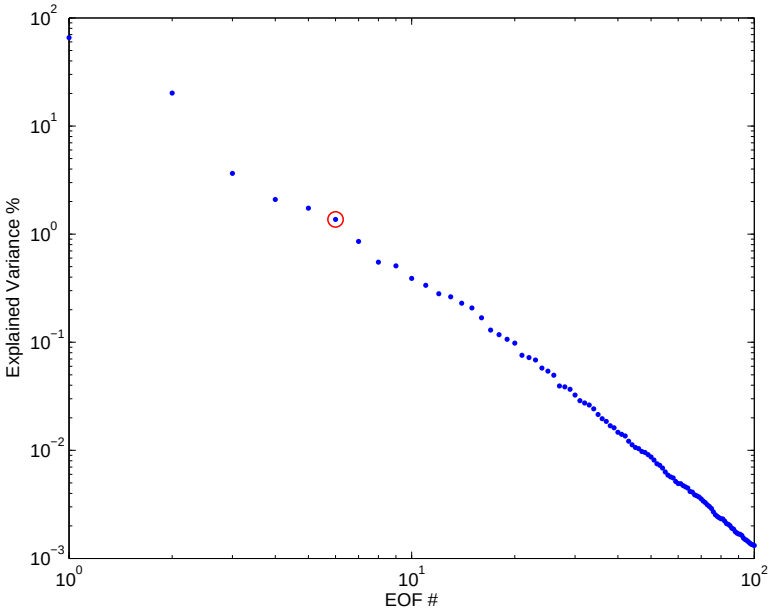


FIGURE 8.7: Variance explained by the first 100 Empirical Orthogonal Functions. The red circle highlights the explained variance of the 6th EOF, the last included in the forecast model.

The first six spatial EOF isolated are then represented in figure 8.8, while the corresponding first six time EOF (for the first week of May 2010) are represented in figure 8.9.

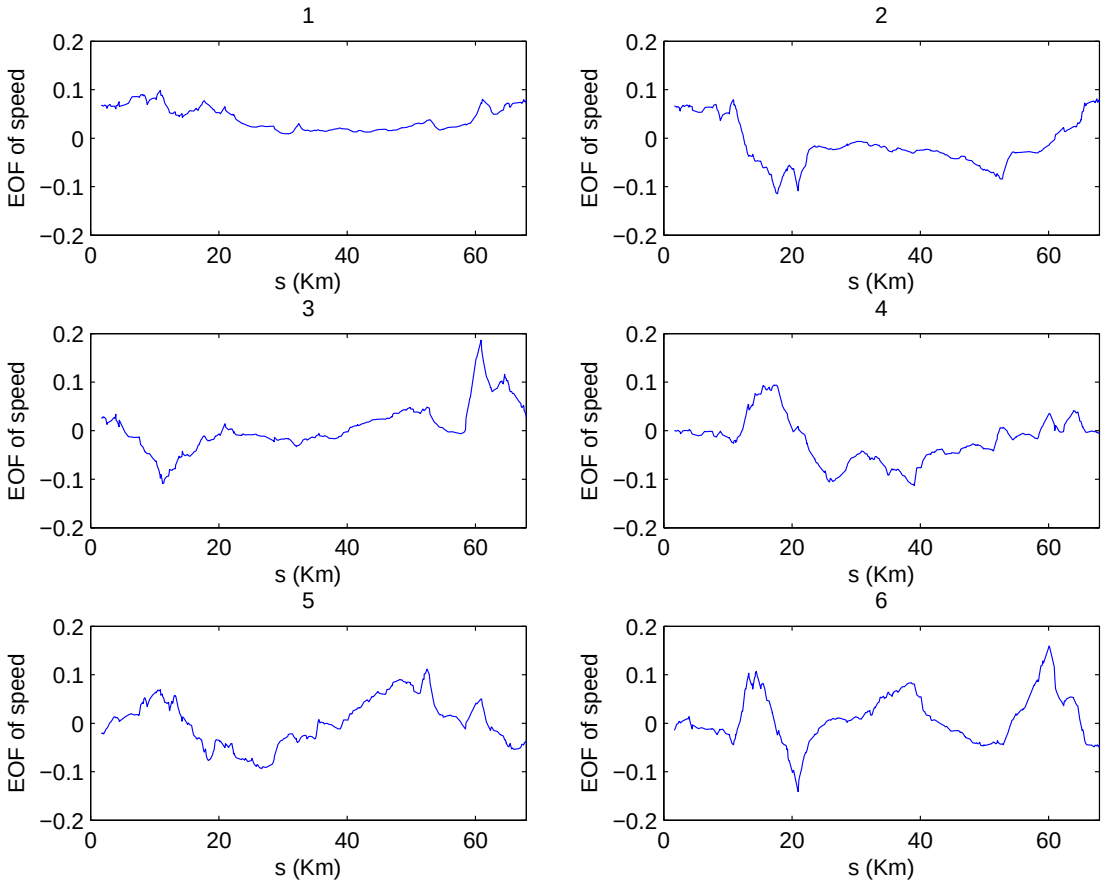


FIGURE 8.8: The first six space EOF.

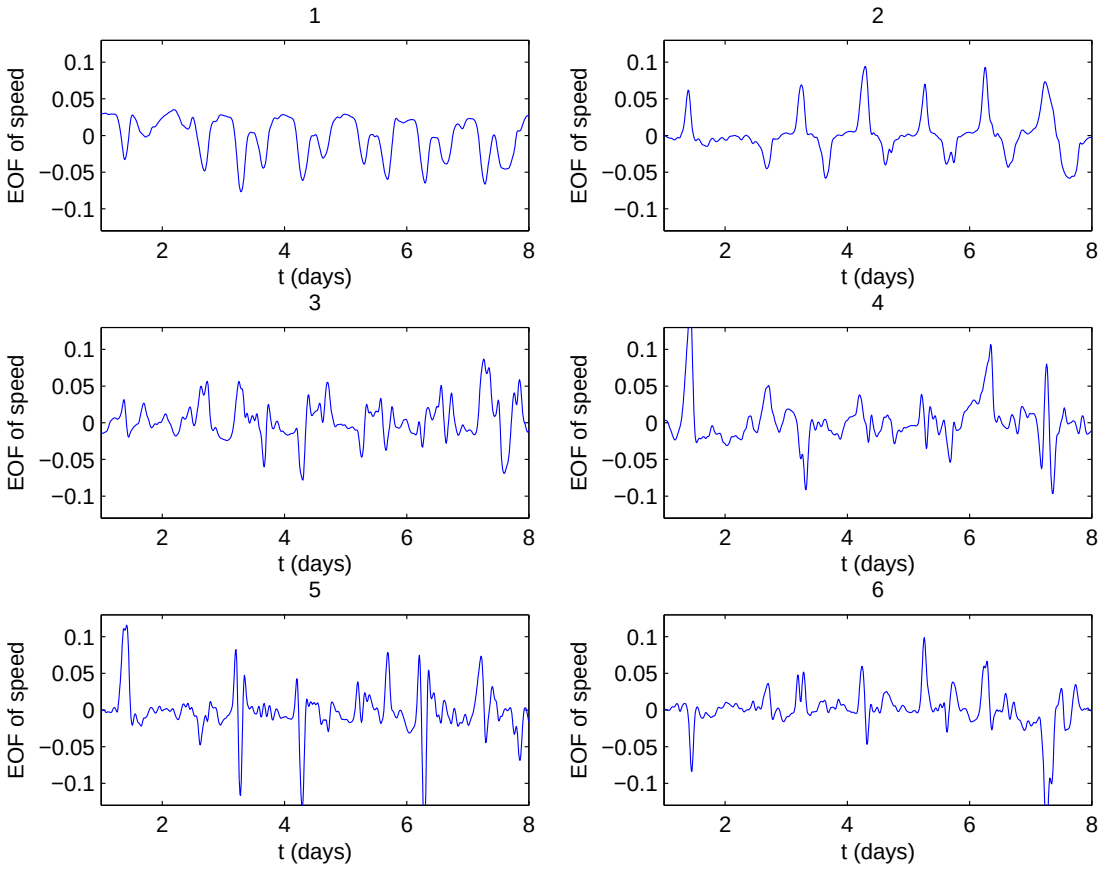


FIGURE 8.9: The first six time EOF.

8.3.2 Predictions

To evaluate the goodness of the statistical prediction formula we compare, for the control sample, the results of the predictions with persistence ($v(t+dt) = v(t)$). For each time t_i, $i > 1$ we can evaluate the forecast value $v_f(t_i)$ and the persistence value $v_p(t_i)$ and compare them with the real data value $v_r(t_i)$. The first quantities we can measure are the mean prediction errors:

- $\mathrm{err}(v)_f^2 = \langle (v_f - v_r)^2 \rangle$
- $\mathrm{err}(v)_p^2 = \langle (v_p - v_r)^2 \rangle$

As the forecast value depends on the choice of the number K of EOF considered, the error err_f is a function of K. On the other hand we expect that the persistence error err_p grows when we demand a farther prediction time dt. In figure 8.10 the values of err_f are plotted in blue while the solid red line is the value of err_p in

our control sample. From this graph appears clearly that the mean error of our forecast is higher than the error of persistence with $dt = 15$ minutes. For $K >> 1$ the difference becomes small, but statistical forecast with this short prediction time remains always worse than persistence.

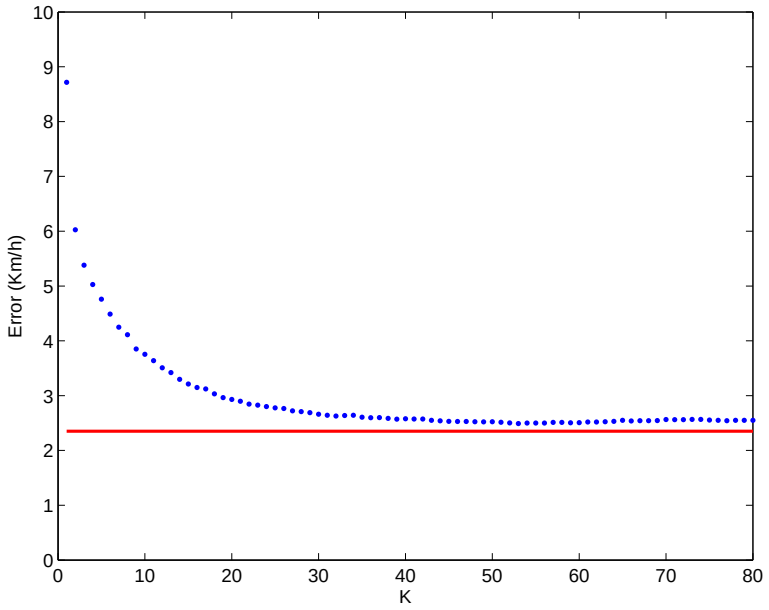


FIGURE 8.10: Mean error in statistical forecasting vs persistence model error

Moreover, we can focus on the variation of the fields between a timeframe and the following, considering the quantities $dv(t_i) = v(t_i) - v(t_{i-1})$. These variations are not taken into account by the persistence model, as the variation of the persistence model is always 0. For the forecast model, instead, there can be correlation between predicted and real variation. This is shown in fig. 8.11 where are represented the values of the correlations between signal variations and the variations estimated from the forecast. An asymptotic behavior for $K > 40$ can be observed also in this graph, and the correlation reaches values of ≈ 0.3. Thus, while errors are comparable with the ones from persistence, the statistical predictor is fairly sensible in predicting sign and relative intensity of the variations.

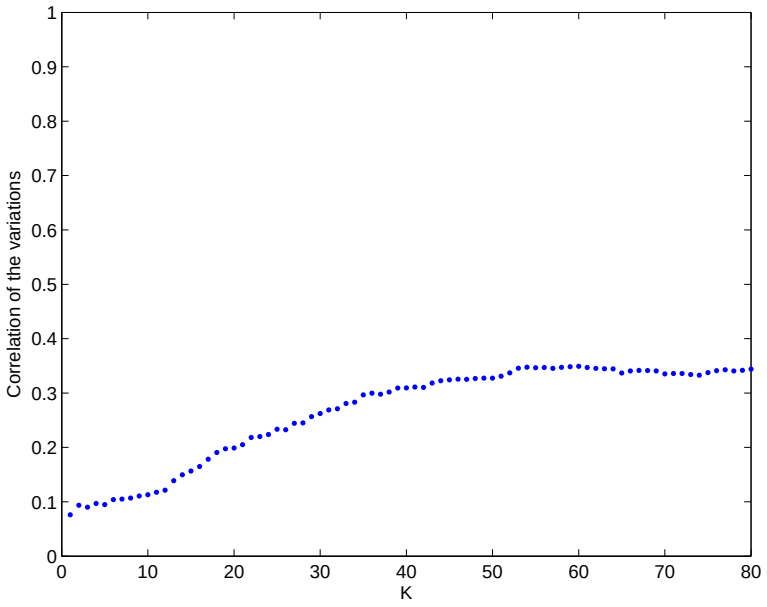


FIGURE 8.11: Correlations of real and forecasted variations for different K.

Furthermore, we can consider covariation instead of correlation, giving in such way a higher weight to high variations and a lower to low ones. The correlation and covariance curves are quite similar, if we exclude the first part. There, the covariance reaches a minimum for $K = 12$ and for $K = 1$ the value is not far from the asymptotic limit for $K \gg 1$. This trend for $K \to 1$ is probably a consequence of the specific optimal covariance approach in the developing of the prediction formula. We can suppose then that the predictor is capable to make good prediction about exceptional wide fluctuations, which are highlighted by covariance, while he makes the most part of the error on the small fluctuation.

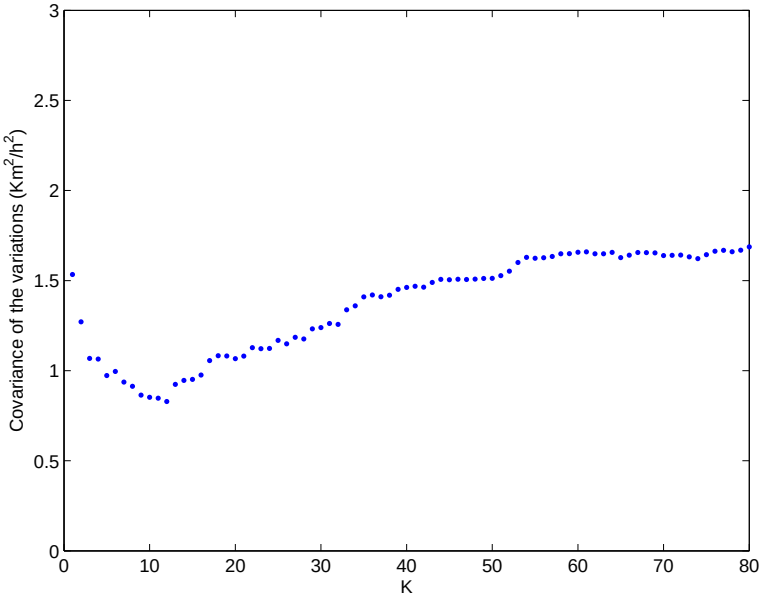


FIGURE 8.12: Covariances of real and forecasted variations for different K.

In conclusion, we remark that even if the statistical forecast gives a mean error higher than persistence at the chosen prediction time of 15 minutes, the formula is able to foresee variations. In particular, comparing covariances and correlations, it appears that strong variations are foreseen better than low ones, at least for low values of K. This is a good result for two reasons. The first is that in our Kalman Filter framework it is important to have a model sensible to variations, especially if quick and strong, because stationary conditions are usually well described by data alone. The second is that, if we want to forecast the state of a whole road network we can easily have a huge number of signals and therefore we need to reduce a lot the dimensionality of the problem. Therefore it will be necessary choosing only a small number of EOF: in our case, keeping for example 8 EOF out of 410, we have already a reduction of a factor 50.

We finally remind that we have two contributes to consider in this forecast model: we have a good contribute in covariance but a bad contribute in errors: it will be necessary to consider both while choosing the correct number K of EOF.

Chapter 9

Conclusions

The focal point in the modeling approach taken in this thesis was finding the key quantities for the description of human mobility, which can then be used as control parameters. Our database of GPS vehicular trajectories guided us in this purpose, and the empirical results are suggesting that one quantity is distinctly predominant among all the others. This quantity is time.

From our analysis has emerged that activities' duration follows a Benford's law, i.e. a power law $p(\tau) \propto 1/t$, modulated by a minor perturbation with a peculiar structure related to working times, which is different from area to area. The scaling exponent -1 has been confirmed by an independent source on visit durations on a web site, but is significantly different from what observed in other datasets of human and animal mobility[16] and it is not consistent with human mobility models developed on the base of phone records[17]. Two alternative interpretations have been proposed to explain the Benford's law of activities' durations, both presuming that circadian rhythm imposes a limit to the time budget that can be spent for the activities of a given day. The first assumes a linear time perception and a progressive assignment of activities durations, limited by the length of the already scheduled tasks. The second assumes a logarithmic time perception: in this case Benford's law represents the limit distribution as far as the support of the distribution is an interval, interval that is dictated by the limit due to circadian rhythm. Outside this limit, a log-normal tail confirms the involvement of a logarithmic time statistics, at least out from the daily schedule.

However, Benford's law is not sufficient to have a full understanding of the role of activity time. In modeling the decisional aspects behind individual mobility there has been a general shift towards its explanation as dynamic on networks.

In this framework, the second ingredient we need is the relationship between the average time spent by an individual in a particular activity $\langle\tau\rangle$ and the number of times k that activity has been performed. This relationship has already been pointed out as crucial for epidemics spreading[71]. In our data, this relationship is described by $\langle\tau\rangle \propto \exp(\gamma k^a)$. This function, and in particular the exponent $a \approx 0.3$, plays a fundamental role in shaping the individual mobility networks, in particular determining the scaling exponent of the degree distribution, and as a consequence how networks grow in time. Furthermore, studying the networks describing the locations visited by the same individuals, has emerged that also the fractional nature of the urbanized territory influences people's use of space.

Benford's law plays a dominant role in the information carried by time mobility patterns, whose analysis has suggested an extremely high predictability of human movements[51]. Analyzing jump patterns where this influence is negligible, the dynamics becomes almost equivalent to a Markov process where the transition probability is given by the empirical weights of the undirected mobility networks. Some fine structure is hidden in this equivalency. It has been possible to identify three different classes of activities: short activities $\tau < 2$h, long activities $2\text{h} < \tau < 12\text{h}$ and ultra long activities $12\text{h} < \tau$. Each of these categories is reasonably homogeneous and contributes to the final structure of the individual mobility patterns. The ultra long activities can be identified with the overnight stop at home, together with only a few other un-frequently taken alternatives. This daily return to home constitutes the base structure of the mobility pattern, which is then developed between one night rest and the following. The short and the long activities are performed in this period of time, but among them exists a hierarchical relationship, where the long activities play a central role and the short activities a subordinate one. Long activities are most likely planned before and play a pivotal role for short activities. In fact, short activities arrange themselves around the long (or ultra long) ones, forming sequences that are repeated in different days. Therefore, contrarily to what would have expected, short activities do not seem to be executed randomly. The homogeneity in the classes indicates that even the shortest ones are eventually able to concur in a systematic mobility, pointed out by the formation of repeated sequences. The nature of the values 2 hours and 12 hours remains an open question. The value of 12 hours seems again to reflect the circadian rhythm, while the value of 2 hours (or, more exactly, 115 minutes) may instead be a duration characteristic of the studied area. The hierarchical structure suggested by our observations is consistent with a recent study[53] proposing that the spatiotemporal structure of human mobility patterns can be

flow-wise partitioned in groups of related nodes called habitats, and that those habitats tend to be more spatially cohesive than the total mobility. The repeated sequences we found may reflect this habitat structure and thus be geographically related. In addition, we found that influences of this division into habitats can be found in the a-posteriori structure of the individual mobility networks, which can be classified as mono- or dipolar according to the characteristics of their hubs.

The fact that activities located one near another are performed in sequence is an effect of the optimization that individuals do on their daily schedule because their mobility is limited by a travel time budget. Travel time budget has been for long time believed to be an universal constant of about 1.2 hours, but our analysis establish that the shape of its distribution in different municipalities is conditioned by economical and census features of the city and by the efficiency of the street network. Travel time budget can be converted in total mobility, i.e. the total length covered in a day, but this conversion is not trivial: for short trips, distance covered d and travel-time t are not simply linked by an average velocity but by a power function: $t \propto d^{\frac{2}{3}}$. Nevertheless, the differences between the activities' time budget partitioning explaining Benford's law and the total mobility partitioning explaining the empirical trip lengths distribution are suggesting that time and space are used and lived in two different ways. Movement are largely planned before they take places, as the whole daily mobility is shared consistently with a simultaneous division of the limited total mobility length, while time is progressively divided considering the activity already performed or planned, and therefore a succession of decisions have to be made during the day.

Travel time is also the quantity we evaluate when choosing between alternative routes. Although, our data indicate that many drivers' choices are non-optimal, as slower routes are chosen when a faster option is available. These mistakes are probably a consequence of incomplete information on traffic conditions. Routes chosen as alternatives have different travel times at the moment of the choice, but the difference is not significant when we take monthly averages. Thus, is probably the average travel-time, known thanks to personal experience, which better represents the cost function taken into account in route choice. In the near future this will probably change, as real time traffic information will be always more integrated in navigation systems. The last chapter of this thesis is about our work in this direction. Data analysis methods originally developed in meteorology have been implemented to deal with one fundamental limit of our data source: data sparseness. Our GPS measures of vehicle trajectories are scattered in space and time. The data aggregation methods we developed solve this problem producing

traffic information represented by continuous fields, which can be used for analysis or control purposes and have been integrated in an info-mobility platform.

This many results will hopefully contribute to the general understanding of human mobility. From our statistical physics perspective, we notice that the observed statistical properties are largely consistent with the maximum entropy principle and the models we have developed are effective in explaining observations without the need of including the interaction of individuals among themselves. Every individual seems to behave as almost independent particle, performing his mobility mostly according to his propensities. Effects due to traffic interactions and collective mobility events appear to be small in average. If mobility could be seen as the realization of many independent individual agenda, its dynamical properties become similar to that of a Boltzmann gas, where drivers organize their mobility by applying a minimization strategy of the interactions with other individuals, like animals that share the same spatial resources[72]. As the system has reached equilibrium, many important aspects such transitory phenomena, which can have strong local effects and great impact on people's life, are hidden by the statistics. In our opinion, is studying these transients that the features of collective mobility dynamics will come to light. For this reason, to achieve a real knowledge of urban mobility systems and their critical states, our efforts have to move from statistical towards dynamical studies.

Appendix A

Progressive time usage model

Let us consider the random choice of k stochastic variables u_i, uniformly distributed in the segment $[0, U_{i-1}]$, while the residual interval is updated with $U_i = U_{i-1} - u_i$. Assuming for sake of simplicity $U_0 = 1$, the variables u_i are given by a combination of independent variables x_j uniformly distributed in the unit segment.

$$\begin{aligned}
u_1 &= x_1 \\
u_2 &= x_2(1 - x_1) \\
u_3 &= x_3(1 - u_1 - u_2) = x_3(1 - x_1)(1 - x_2) \\
&\vdots \\
u_k &= x_k(1 - x_1) \dots (1 - x_{k-1})
\end{aligned}$$

We note that if x_j is uniformly distributed in $[0, 1]$ we get the same for $(1 - x_j)$.

Finding the distribution of u_i is equivalent to finding the distribution of:

$$y_k = x_1 x_2 \dots x_k$$

We proceed by induction on k starting from $k = 2$

$$y_2 = x_1 x_2$$

We may calculate

$$p(y_2) = \int_0^1 \int_0^1 \delta(x_1 x_2 - y_2) dx_1 dx_2$$

with the change of variables

$$\begin{aligned} u &= x_2 \\ v &= x_1 x_2 \end{aligned} \Rightarrow dudv = \begin{vmatrix} 0 & 1 \\ x_2 & x_1 \end{vmatrix} dx_1 dx_2$$

$$dx_1 dx_2 = \frac{dudv}{u} \qquad \begin{aligned} u &\in [0,1] \\ v &\in [0,u] \end{aligned}$$

obtaining:

$$p(y_2) = \int_0^1 \frac{du}{u} \int_0^u \delta(v - y_2) dv = \int_0^1 \delta(v - y_2) dv \int_v^1 \frac{du}{u} = -\log y_2$$

Assuming that as true for k we may determine $p(y_{k+1})$ where $y_{k+1} = x_{k+1} y_k$:

$$p(y_{k+1}) = \int_0^1 \int_0^1 \rho(y_k) \delta(x_{k+1} y_k - y_{k+1}) dx_{k+1} dy_k$$

With:

$$\begin{aligned} u &= x_{k+1} \\ v &= x_{k+1} y_k \end{aligned} \Rightarrow dudv = \begin{vmatrix} 0 & 1 \\ x_{k+1} & y_k \end{vmatrix} dx_{k+1} dy_k$$

We find:

$$p(y_{k+1}) = \int_0^1 \frac{du}{u} \int_0^u \rho\left(\frac{v}{u}\right) \delta(v - y_{k+1}) dv = \int_0^1 \delta(v - y_{k+1}) dv \int_v^1 \rho\left(\frac{v}{u}\right) \frac{du}{u}$$

The explicit calculation gives:

$$\int_v^1 \rho\left(\frac{v}{u}\right) \frac{du}{u} = \int_v^1 \frac{(-\log(v/u))^{k-1}}{(k-1)!} \frac{du}{u} = \int_0^{-\log v} \frac{z^{k-1}}{(k-1)!} dz = \frac{(-\log v)^k}{k!}$$

Where we have introduced the variable $z = -\log(\frac{v}{u})$.
Finally, we obtain:

$$p(y_{k+1}) = \int_0^1 \delta(v - y_{k+1}) \frac{(-\log v)^k}{k!} dv = \frac{(-\log y_{k+1})^k}{k!}$$

and therefore u_k follows the distribution:

$$p(u_k) \sim \frac{(-\log u_k)^{k-1}}{(k-1)!}$$

Let u be any u_k, where all $k \in [1, n]$ has equal probability of being represented. The distribution of u reads:

$$p(u) \sim C_n \sum_{k=1}^{n} \frac{(-\log u_k)^{k-1}}{(k-1)!} \simeq C_n e^{-\ln u} = \frac{C_n}{u}$$

Where C_n is a normalization factor.

Appendix B

Random Partitioning of the Total Mobility

It is possible to compute in analytical way the single trip length distribution, as the distribution realized by uniformly spreading k points into a given segment of length L. A simple calculation provides the single trip length distribution in the form

$$p_{N,L}(x) = \frac{c}{L}\sum_{k=1}^{N}(k+1)ka^k(1-x/L)^{k-1} \tag{B.1}$$

where c is a normalizing factor and N is the maximum number of daily activities; we remark that the choice of the points in the segment is contextual without any time-ordering. It is quite natural to assume that there should exist a correlation between the number of daily activities N and the daily mobility length L, but the GPS data do suggest that this correlation is weak.

Let us consider k stochastic variables uniformly distributed in the unit segment, the probability that a segment of length $\leq x$ is empty can be estimated according

$$\mathcal{P}(\leq x) = 1 - (1-x)^k$$

As a consequence the probability density that a certain segment x is empty is given by

$$p(x) = \frac{d\mathcal{P}}{dx} = k(1-x)^{k-1}$$

Therefore if one choices randomly an integer number k in the interval $[1, N]$, the probability density for a segment of length x conditioned by the choice k is

$$p_k(x) \propto (k+1)k(1-x)^{k-1} \qquad x \in [0,1] \tag{B.2}$$

since we have to take into account $k+1$ possible segments. The probability (B.2) has to be weighted by the probability $p(k) \propto a^k$ to have k points so that the probability density to detect a segment of length x for any choice k is

$$p_N(x) = \frac{(1-a)^2}{(2-a)a(1-a^N) - Na^{N+1}(1-a)} \sum_{k=1}^{N} (k+1)ka^k(1-x)^{n-1} \tag{B.3}$$

where we have introduced a normalizing factor.

Appendix C

Link between degree distribution and universal activity time

There is a strict relation between the activity degree distribution (see fig. 5.1 in this dissertation) and the existence of an universal distribution probability $f(u)$ in eq.(3.5). Indeed, taking advantage from the dependence of $\langle t \rangle_k$ on the degree k pointed out by experimental observations (see fig. 3.4) we perform the change of variables

$$\begin{cases} t & = t \\ u & = t/\langle t \rangle_k \end{cases} \tag{C.1}$$

in the join probability distribution $p(k,t)$ of degree and downtime (cfr. eq.(5.1)). Using the definition (3.3), we get the new distribution

$$p'(u,t) = f(u)k(u)p(k(u))\frac{dk}{du} = -f(u)kp(k)\frac{\langle t \rangle_k}{t}\,(d\langle t \rangle_k over dk)^{-1}$$

where k has to be read $k(u)$ in the r.h.s. and $p(k)$ is the activity degree distribution. In the previous formula we approximate interpolate the discrete variable k with a continuous variable. By integrating of u we have to recover the Benford's law $\propto 1/t$ for the global activity downtime distribution (see fig. 3.1 left). Since $f(u)$ is normalized as probability distribution, this is possible if

$$kp(k)\langle t \rangle_k \left(\frac{d\langle t \rangle_k}{dk}\right)^{-1} = \text{const.} \tag{C.2}$$

According to the interpolation $< t >_k \propto \exp \gamma k^a$ of the experimental data as shown in the figure 3.4, we explicitly have

$$\frac{d\langle t\rangle_k}{dk} \propto k^{a-1} e^{\gamma k^a} \propto k^{a-1}\langle t\rangle_k$$

therefore the condition (C.2) reads

$$k^{2-a} p(k) = \text{const.}$$

i.e. a power law distribution of the activity degree with exponent ≤ 2. This is consistent with the experimental observations as shown by the figure 5.1.

Appendix D

Quantitative Analysis on Individual Mobility Networks

In this appendix we want to study the networks described by the origin-destination mobility of people. In particular we will analyze a GPS dataset where the movements of 32457 vehicles has been recorded for a month (March 2008) in the Province of Florence.

The main intent of this study is to isolate a set of fundamental observables, which could permit the classification of different species of moving individuals.

D.1 Individual Mobility Networks

In our data each vehicle makes a series of trips and each trip has a point of origin and one of destination. Through a gravitational clustering process, executed separately for each different vehicle, these points have been assigned to a specific node that identifies a circular area with diameter 400m. Each trip is then associated to a directed link between the origin node and the destination node. This way we have built directed multigraphs, but if we assign as the weight w_{ij} the number of links from the node i to the node j, we can identify $A_{d,w} = \{w_{ij}\}$ as the weighted adjacency matrix of a Directed Weighted Network. We can then easily extract from $A_{d,w}$:

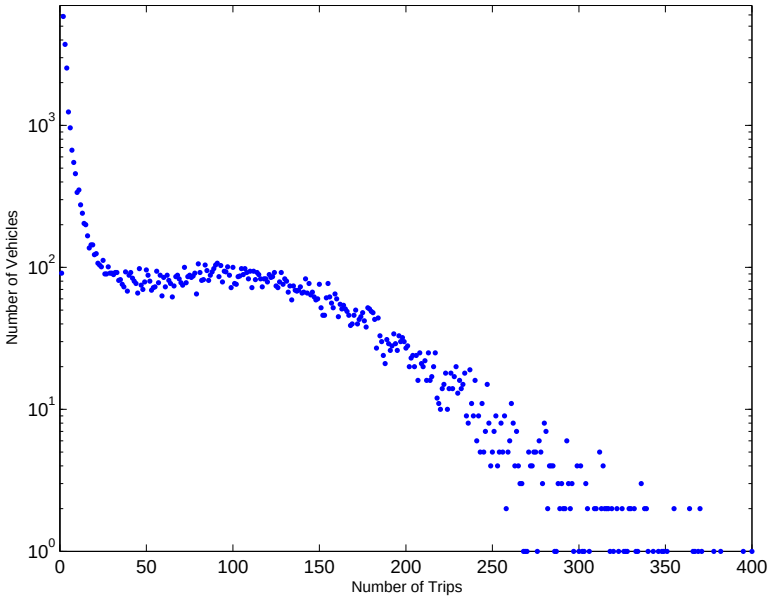
- Undirected Weighted Network
  $A_{u,w} = \{w^u_{ij}\} = A_{d,w} + A^{\mathrm{T}}_{d,w}$

- Directed Unweighted Network
  $A_{d,u} = \{a_{ij}\} \quad : \quad a_{ij} = 1 \text{ if } w_{ij} > 0 \quad , \quad a_{ij} = 0 \text{ if } w_{ij} = 0$
- Undirected Unweighted Network
  $A_{u,u} = \{a^u_{ij}\} \quad : \quad a^u_{ij} = 1 \text{ if } (w^u_{ij}) > 0 \quad , \quad a^u_{ij} = 0 \text{ if } w^u_{ij} = 0$

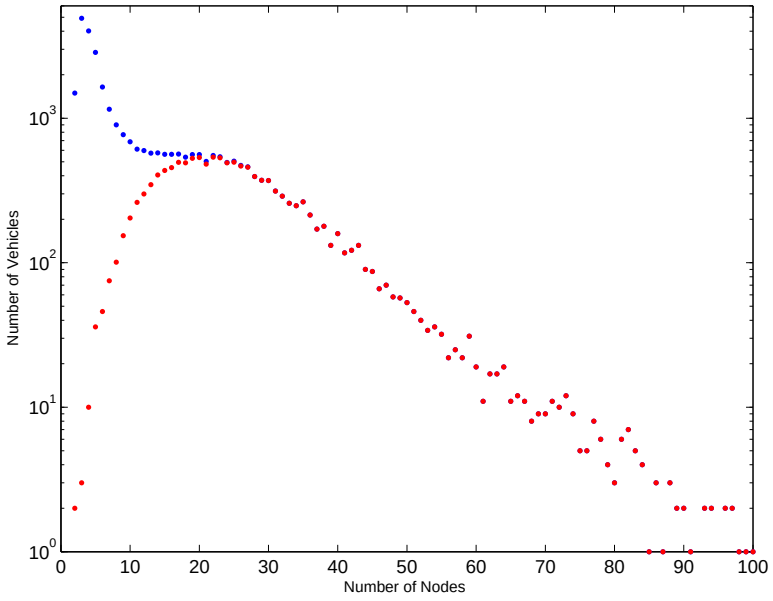
Assuming that the same person always drives each vehicle, we can therefore associate an Individual Mobility Network to each sequence of data relative to one of our vehicles.

D.2 Data Filtering

We have in our dataset over 30000 networks, but a large part represents the mobility of individuals not living in the area in analysis, visitors who have made only a few trips an visiting few nodes. Observing the number of trips distribution:



we have decided to consider as a sample for our study just the 13127 vehicles which have maid at least 30 trips in the 31 days considered in this study. As we can see considering the Number of Nodes distribution:



it appears that we have made a good selection of the networks with a significant number of nodes (red) in the whole sample (blue).

It is important to highlight here that we have not, as usual in network studies, a huge ($N \gg 1$) network but an ensemble of small networks ($N < 100$). We can compute then a series of observables relative to each single network or we can evaluate others observable relative every node or alternatively some nodes in particular.

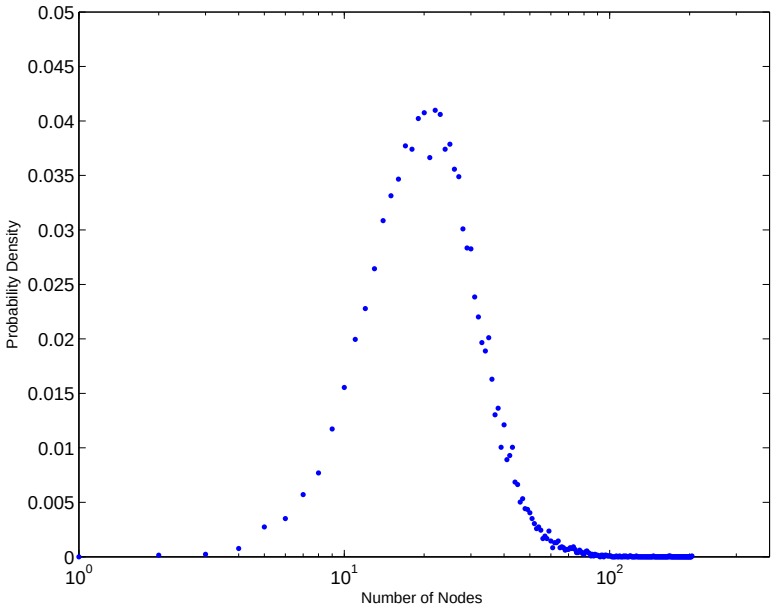
D.3 Classical Observables

Our work on this dataset has been the computation, for each network, of some classical observables of the network theory. These observables are usually evaluated for simple (i.e. undirected and unweighted) networks. Most networks' observables are properties of single links or nodes (local parameters), in that case we have to extract a global parameter, as the mean, to represent the whole network. Many of these observables have been computed using the Matlab Boost Graph Library.

D.3.1 Network Size

The number N of nodes in each network represents the number of different sites visited by each individual during the month. His logarithm is the upper bound

to the information entropy of the sequence of sites visited by that individual. Therefore we can choose $\log(N)$ as a natural measure of the size of the network.



D.3.2 Connectivity Degree

The Connectivity Degree of a node (often just called degree of a node) is the number of links entering to (IN-degree) or exiting from (OUT-degree) a given node. It is easily defined from the adjacency matrix as:

$$k_{in}(i) = \sum_j w_{ij}$$

$$k_{out}(j) = \sum_i w_{ij}$$

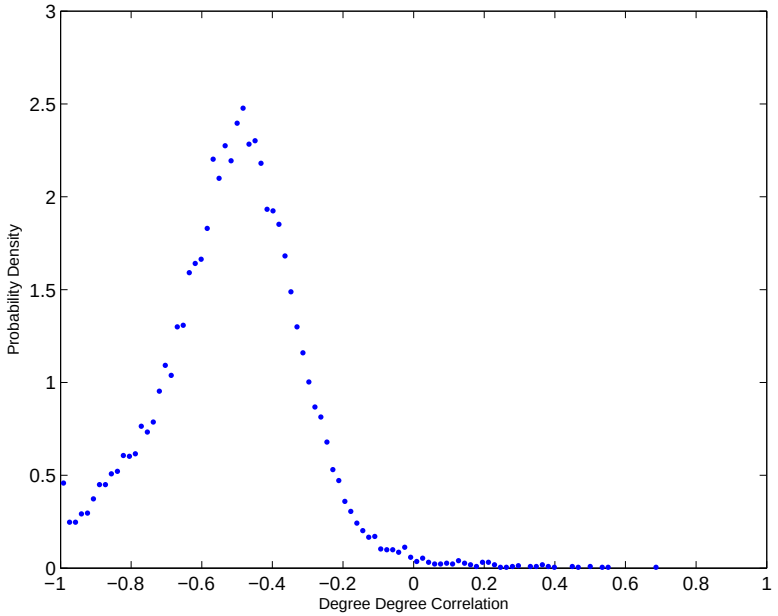
For undirected networks one can speak just of degree as $k(i) = k_{in}(i) = k_{out}(i)$. For our weighted and directed mobility networks also we can simply speak of degree k, because, if there are no losses in the signal, each time a trips ends in a node, the next trip will start from that node.

In the following with k we will identify the weighted degree while we will use tilde notation $\tilde{k}$ for the unweighted degree.

For all the nodes of a network we can estimate the correlation between the degree of the node $\tilde{k}(i)$ and the mean of the degrees of his neighbors $\tilde{k}_n(i)$.

$$DD = \mathrm{corr}(\tilde{k}, \tilde{k}_n)$$

As we see:

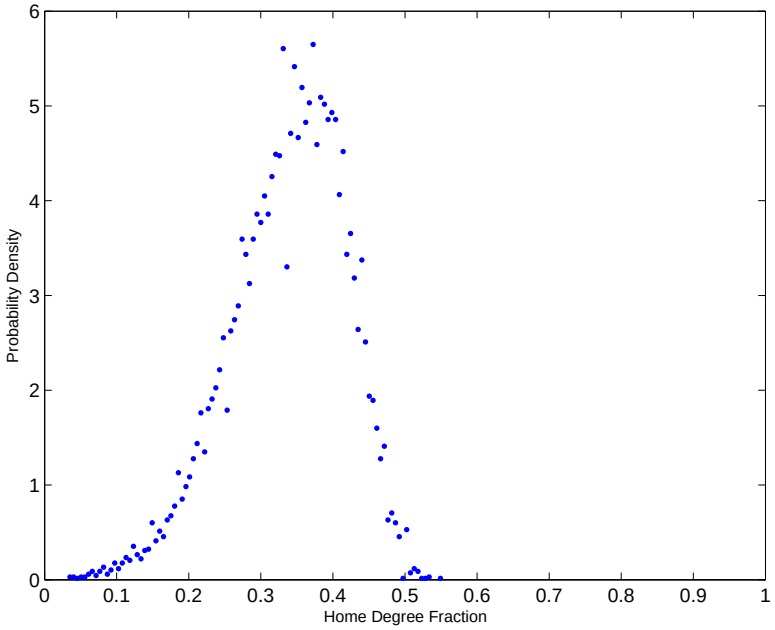


$\tilde{k}(i)$ and $\tilde{k}_n(i)$ are anti-correlated, showing the disassortativity of the mobility networks.

The values of the degree-degree correlation for the weighted and unweighted form of network describing the same individual mobility are highly correlated (0.87): we have then chosen to show in the figure just the unweighted value.

We associate to the home of a particular individual his most visited node, i.e. the node with the greater weighted degree. Then we will call home degree fraction k_h the fraction of connectivity taken by this main hub.

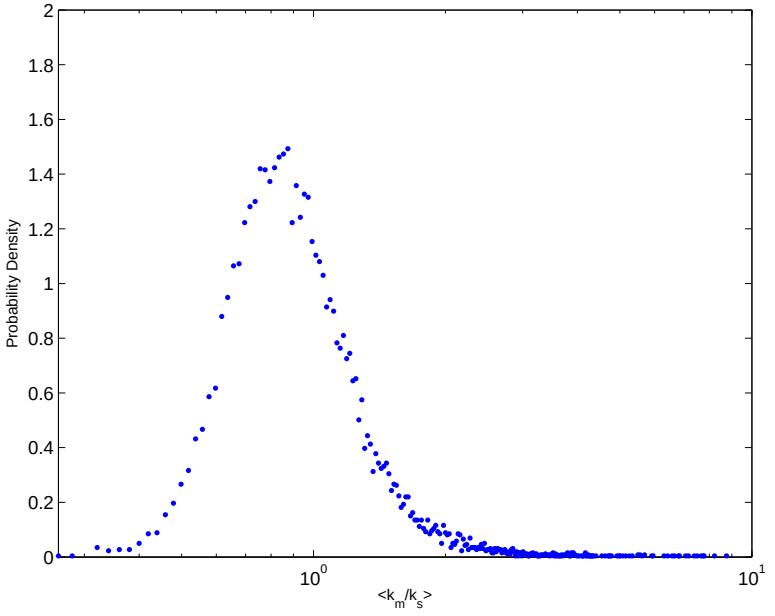
$$k_h = \frac{k(h)}{\sum_i k(i)}$$



The weighted degree k is greater than the unweighted degree $\tilde{k}$. We want to evaluate the tendency to repeat trips by estimating:

$$R_{wu} = \left\langle \frac{k(i)}{\tilde{k}(i)} \right\rangle_i$$

We can see that the distribution of $\log(R_{wu})$ is bell shaped and symmetrical, then we will utilize this logarithm as a measure of the repetition of the links.

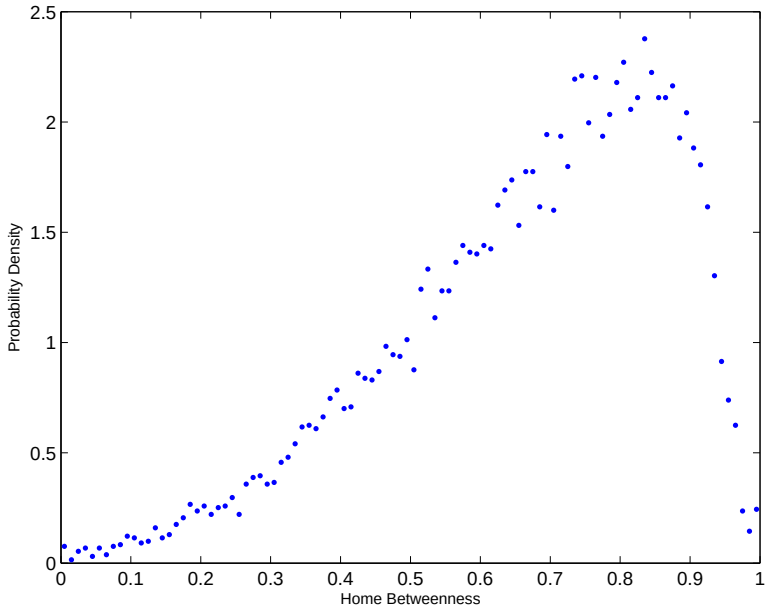


D.3.3 Betweenness Centrality

Using an algorithm like the Dijkstra it is possible to find all the shortest paths P_{mn} between the nodes m and n. The Normalized Betweenness Centrality is then defined for each node or link as the fraction of these shortest paths passing through that node/link:

$$BC(i) = \frac{1}{BC_{max}} \frac{\sum_{m,n} P_{mn}(i)}{\sum_{m,n} P_{mn}}$$

where $BC_{max} = \frac{(N-1)(N-2)}{2}$ is the total number of shortest paths in the network.



In our dataset we find that the 90% of the homes have also the top betweenness centrality. We do not show much difference considering weighted or unweighted networks (the two results are 0.84 correlated).

D.3.4 Clustering coefficient

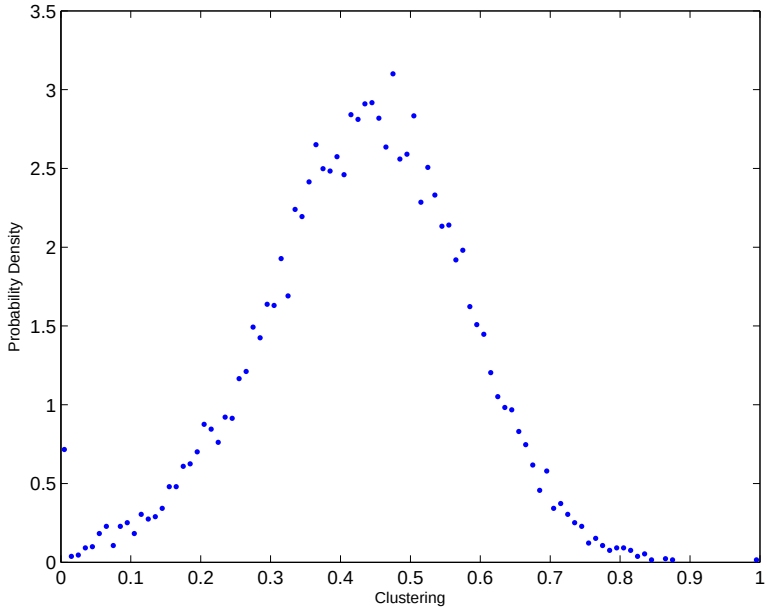
The clustering coefficient of a node is the measure of the fraction of neighbors that are connected between them. It's value for the node i is:

$$C_i = \frac{2e_i}{\tilde{k}_i(\tilde{k}_i - 1)}$$

where e_i is the number of connections between first neighbors.

For the whole network is defined the clustering coefficient as $C = \langle C_i \rangle_i$.

For each network we can evaluate C and their distribution is shown in figure:



D.3.5 Correlations

For the intent of isolating an effective set of features for the developing of a clustering procedure on these Individual Networks is worth to evaluate if the observables are correlated.

	k_h	BC(h)	DD	C	log(N)	log(R_{wu})
k_h	1	0.61	-0.54	0.18	-0.53	0.61
BC(h)	0.61	1	-0.12	-0.05	-0.02	0.10
DD	-0.54	-0.12	1	-0.47	0.66	-0.63
C	0.18	-0.05	-0.47	1	-0.16	0.29
log(N)	-0.53	-0.02	0.66	-0.16	1	-0.64
log(R_{wu})	0.61	0.10	-0.63	0.29	-0.64	1

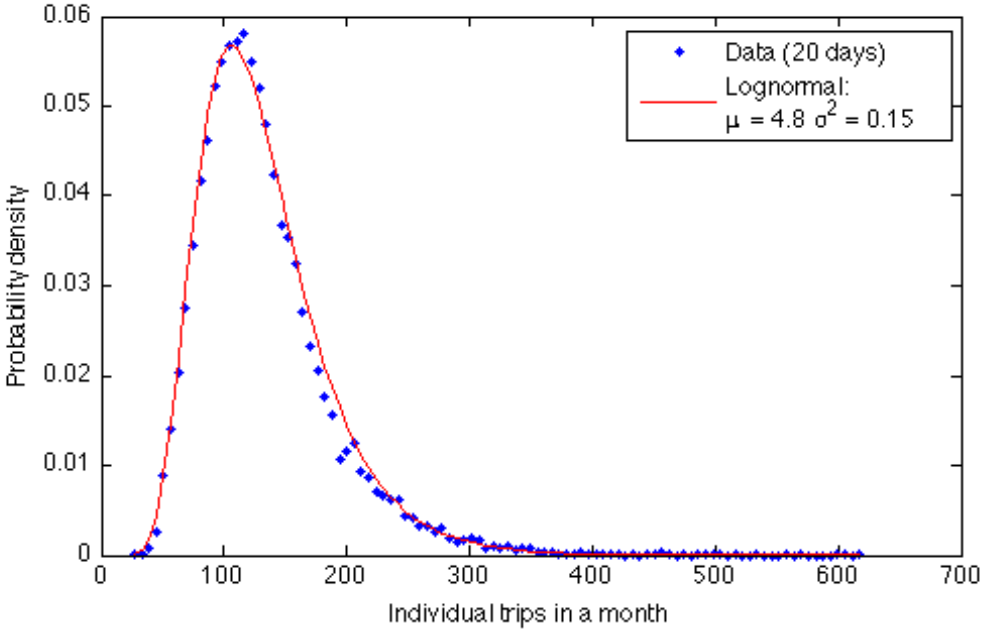
Appendix E

Emilia-Romagna Trip and Node number distributions

E.1 Trips

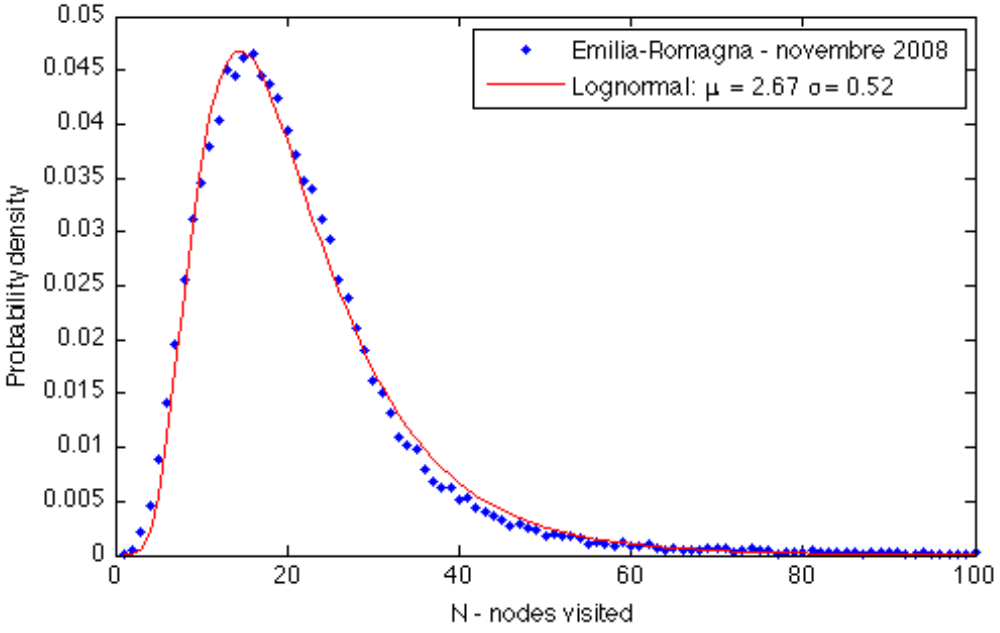
The total number of trips T made by an individual depends on the number of days of mobility. We have isolated all drivers who performed mobility in all the 20 working days included in the Emilia-Romagna dataset, and for these subset we have that T is log-normally distributed:

$$p(T) = \frac{1}{\sqrt{2\pi}\sigma_T T} \exp\left(-\frac{(\log T - \mu_T)^2}{2\sigma_T^2}\right) \tag{E.1}$$



E.2 Nodes

In every day of mobility each individual visits different places. However, many of those places have already been visited before, even very recently. We have considered the number of different places visited N by each individual with 20 days in mobility. We may notice that in this subset that N too is distributed log-normally:



$$p(N) = \frac{1}{\sqrt{2\pi}\sigma_N N} \exp\left(-\frac{(\log N - \mu_N)^2}{2\sigma_N^2}\right)$$

A similar curve can be deduces also from the random entropy curve [51].

Appendix F

Discrete Linear Kalman Filter

The Discrete Linear Kalman Filter, which is based on the stochastic-dynamic system[1]:

$$\mathbf{x}_{k+1} = \phi_k \mathbf{x}_k + \mathbf{w}_k \tag{F.1}$$

$$\mathbf{z}_{k+1} = \mathbf{H}_k \mathbf{x}_k + \mathbf{v}_k \tag{F.2}$$

where x_k is an n-vector the model state (the estimate) at time t_k while z_k is an m-vector representing the observed state (the empirical information). The evolution of the model state is regulated by the $n \times n$ transition matrix ϕ_k while observed state and model state are related by the $m \times n$ observation operator $\mathbf{H}_k$. Both model and observation are subject to errors: w_k is the model error and v_k is the observation error. If we had the true state x and true measurements z, given our imperfect discrete forecast model and observation operator the state would evolve according to the equations (F.1)(F.2).

Let us make the assumption that those errors are white, unbiased and independent of each other:

$$\langle \mathbf{w}_k \rangle = 0 \tag{F.3}$$

$$\langle \mathbf{w}_k \cdot (\mathbf{w}_l)^t \rangle = \mathbf{Q}\delta_l^k \tag{F.4}$$

$$\langle \mathbf{v}_k \rangle = 0 \tag{F.5}$$

$$\langle \mathbf{v}_k \cdot (\mathbf{v}_l)^t \rangle = \mathbf{R}\delta_l^k \tag{F.6}$$

$$\langle \mathbf{w}_k \cdot (\mathbf{v}_l)^t \rangle = 0 \tag{F.7}$$

[1] This appendix is completely based on Saroja Polavarapu lecture notes[73]

The Kalman Filter problem is this: given a prior (background) estimate x_k^f, of the system state at time t_k, what is the update or analysis x_k^a, based on the measurements z_k? The background x_k^f bears a superscript f referring to the fact that it is derived from a model forecast. The superscript a refers to the analysis, or estimate. At time t_{k+1} a forecasting of the actual state is made from the preceding state:

$$\mathbf{x}_{k+1}^f = \phi_k \mathbf{x}_k^a \tag{F.8}$$

$$\mathbf{P}_{k+1}^f = \phi_k \mathbf{P}_k^a \phi_k^t + \mathbf{Q}_k \tag{F.9}$$

Then, by requesting an unbiased estimate with minimal analysis error variance, the new estimate (or analysis) is defined in the linear, recursive form:

$$\mathbf{x}_k^a = \mathbf{x}_k^f + \mathbf{K}_k(\mathbf{z}_k - \mathbf{H}_k \mathbf{x}_k^f) \tag{F.10}$$

$$\mathbf{P}_k^a = (\mathbf{I} - \mathbf{K}_k \mathbf{H}_k)\mathbf{P}_k^f \tag{F.11}$$

where $\mathbf{K}_k$ is the Kalman gain:

$$\mathbf{K}_k = \mathbf{P}_k^f \mathbf{H}_k^t (\mathbf{H}_k \mathbf{P}_k^f \mathbf{H}_k^t + \mathbf{R}_k)^{-1} \tag{F.12}$$

Appendix G

Asymptotic limits for the Kalman Equations

G.1 Perfect Model: $\sigma_q^2 = 0$

For a single variable only the variance Kalman Equations are:

$$K_k = \frac{p_k^f}{\sigma_o^2 + p_k^f} \tag{G.1}$$

$$p_k^f = \alpha_k p_{k-1}^a \tag{G.2}$$

$$p_k^a = \frac{\sigma_o^2 p_k^f}{\sigma_o^2 + p_k^f} \tag{G.3}$$

where α_k represents the linearization of the model ϕ at time k.

Given these equations in [74] it is demonstrated that, for $\alpha_k = \alpha > 1$:

$$\lim_{k\to\infty} p_k^f = \tilde{p}^f = \sigma_o^2(\alpha - 1) \tag{G.4}$$

$$\lim_{k\to\infty} p^a = \sigma_o^2 \frac{\alpha - 1}{\alpha} \tag{G.5}$$

$$\lim_{k\to\infty} K = \frac{\alpha - 1}{\alpha} \tag{G.6}$$

The three limits actually exists also for $\alpha = 1$, but are all 0.

These limits have been exactly calculated. It is possible to estimate them also looking for stable stationary points of the respective successions. For example at

each step p_k^f evolves as:

$$p_k^f = \frac{\alpha\sigma_o^2 p_{k-1}^f}{\sigma_o^2 + p_{k-1}^f}$$

And the limit $\to \infty$ correspond to the stationary point $\tilde{p}^f$:

$$\begin{aligned}
\tilde{p}^f &= \frac{\alpha\sigma_o^2\tilde{p}^f}{\sigma_o^2 + \tilde{p}^f} \\
\sigma_o^2\tilde{p}^f + \tilde{p}^{f2} &= \alpha\sigma_o^2\tilde{p}^f \\
\tilde{p}^f\sigma_o^2(1-\alpha) + \tilde{p}^{f2} &= 0 \\
\tilde{p}^f &= \sigma_o^2(\alpha - 1)
\end{aligned}$$

If we want to check the stability of $\tilde{p}^f$ we can perturb the stationary point $\tilde{p}^f$ with an $\epsilon_{k-1} \ll 1$. The following value in the succession will be:

$$\begin{aligned}
\tilde{p}^f + \epsilon_k &= \frac{\alpha\sigma_o^2(\tilde{p}^f + \epsilon_{k-1})}{\sigma_o^2 + \tilde{p}^f + \epsilon_{k-1}} \\
&= \frac{\alpha\sigma_o^2\tilde{p}^f}{\sigma_o^2 + \tilde{p}^f + \epsilon_{k-1}} + \frac{\alpha\sigma_o^2\epsilon_{k-1}}{\sigma_o^2 + \tilde{p}^f + \epsilon_{k-1}} \\
&= \frac{\frac{\alpha\sigma_o^2\tilde{p}^f}{\sigma_o^2+\tilde{p}^f}}{1 + \frac{\epsilon_{k-1}}{\sigma_o^2+\tilde{p}^f}} + \frac{\frac{\alpha\sigma_o^2\epsilon_{k-1}}{\sigma_o^2+\tilde{p}^f}}{1 + \frac{\epsilon_{k-1}}{\sigma_o^2+\tilde{p}^f}} \\
&\simeq \frac{\alpha\sigma_o^2\tilde{p}^f}{\sigma_o^2 + \tilde{p}^f}\left(1 - \frac{\epsilon_{k-1}}{\sigma_o^2 + \tilde{p}^f}\right) + \frac{\alpha\sigma_o^2\epsilon_{k-1}}{\sigma_o^2 + \tilde{p}^f}\left(1 - \frac{\epsilon_{k-1}}{\sigma_o^2 + \tilde{p}^f}\right) \\
&\simeq \left(\frac{\alpha\sigma_o^2\tilde{p}^f}{\sigma_o^2 + \tilde{p}^f}\right) + \epsilon_{k-1}\alpha\frac{\sigma_o^2}{\sigma_o^2 + \tilde{p}^f}\left(1 - \frac{\tilde{p}^f}{\sigma_o^2 + \tilde{p}^f}\right) \\
&= \tilde{p}^f + \epsilon_{k-1}\alpha\left(\frac{\sigma_o^2}{\sigma_o^2 + \tilde{p}^f}\right)^2 \\
&= \tilde{p}^f + \epsilon_{k-1}\alpha\left(\frac{\sigma_o^2}{\sigma_o^2 + (\sigma_o^2(\alpha-1))}\right)^2 \\
&= \tilde{p}^f + \epsilon_{k-1}\frac{1}{\alpha}
\end{aligned}$$

And the trivial error succession $\epsilon_k = \epsilon_{k-1}\alpha$ leads to the convergence condition $\alpha > 1$.

G.2 Imperfect Model: $\sigma_q^2 > 0$

Let the model be constant in time ($\alpha_k = \alpha$) and the predictions errors have a variance $\sigma_q^2 > 0$. The Kalman equations now read:

$$K_k = \frac{p_k^f}{\sigma_o^2 + p_k^f} \tag{G.7}$$

$$p_k^f = \alpha p_{k-1}^a + \sigma_q^2 \tag{G.8}$$

$$p_k^a = \frac{\sigma_o^2 p_k^f}{\sigma_o^2 + p_k^f} \tag{G.9}$$

We want now to find and analyze the stationary points of p^f and p^a, verifying that they are the limits of the successions.

G.2.1 p^f

At each step the p_k^f evolves as:

$$p_k^f = \frac{\alpha\sigma_o^2 p_{k-1}^f}{\sigma_o^2 + p_{k-1}^f} + \sigma_q^2 \tag{G.10}$$

A priori in this case there can be two the stationary points $\tilde{p}_\pm^f$:

$$\begin{aligned}
&\tilde{p}^f = \frac{\alpha\sigma_o^2\tilde{p}^f}{\sigma_o^2 + \tilde{p}^f} + \sigma_q^2 \\
&(\sigma_o^2 + \tilde{p}^f)\tilde{p}^f = \alpha\sigma_o^2\tilde{p}^f + (\sigma_o^2 + \tilde{p}^f)\sigma_q^2 \\
&\tilde{p}^{f2} + ((1-\alpha)\sigma_o^2 - \sigma_q^2)\tilde{p}^f - \sigma_o^2\sigma_q^2 = 0 \\
&\tilde{p}_\pm^f = \frac{1}{2}\left(-((1-\alpha)\sigma_o^2 - \sigma_q^2) \pm \sqrt{((1-\alpha)\sigma_o^2 - \sigma_q^2)^2 + 4\sigma_o^2\sigma_q^2}\right)
\end{aligned}$$

But the one-dimensional case $\tilde{p}^f$ is simply a variance and should then be > 0. Therefore, calling $B = -((1-\alpha)\sigma_o^2 - \sigma_q^2)$, we have the condition:

$$B \pm \sqrt{B^2 + 4\sigma_o^2\sigma_q^2} > 0$$

The + equation:

$$\sqrt{B^2 + 4\sigma_o^2\sigma_q^2} > -B$$

is always satisfied, while the $-$ is never satisfied:

$$\sqrt{B^2 + 4\sigma_o^2\sigma_q^2} < B$$

Thus the $\tilde{p}^f_-$ solution has to be excluded and the stationary point is only:

$$\tilde{p}^f = \frac{1}{2}\left((\sigma_q^2 + (\alpha-1)\sigma_o^2) + \sqrt{((1-\alpha)\sigma_o^2 - \sigma_q^2)^2 + 4\sigma_o^2\sigma_q^2}\right) \tag{G.11}$$

In the $\alpha = 1$ case in particular we will have as stationary point:

$$\tilde{p}^f_{\alpha=1} = \frac{1}{2}\left(\sigma_q^2 + \sqrt{\sigma_q^4 + 4\sigma_o^2\sigma_q^2}\right) \tag{G.12}$$

G.2.2 p^a

p^a_k evolves as:

$$p^a_k = \frac{\sigma_o^2(\alpha p^a_{k-1} + \sigma_q^2)}{\sigma_o^2 + (\alpha p^a_{k-1} + \sigma_q^2)} \tag{G.13}$$

The stationary points $\tilde{p}^a$ are the solutions of:

$$\begin{aligned}
&\tilde{p}^a = \frac{\sigma_o^2(\alpha\tilde{p}^a + \sigma_q^2)}{\sigma_o^2 + (\alpha\tilde{p}^a + \sigma_q^2)} \\
&(\sigma_o^2 + \alpha\tilde{p}^a + \sigma_q^2)\tilde{p}^a = \sigma_o^2(\alpha\tilde{p}^a + \sigma_q^2) \\
&\alpha\tilde{p}^{a2} + ((1-\alpha)\sigma_o^2 + \sigma_q^2)\tilde{p}^a - \sigma_o^2\sigma_q^2 = 0 \\
&\tilde{p}^a_\pm = \frac{1}{2\alpha}\left(-((1-\alpha)\sigma_o^2 + \sigma_q^2) \pm \sqrt{((1-\alpha)\sigma_o^2 + \sigma_q^2)^2 + 4\alpha\sigma_o^2\sigma_q^2}\right)
\end{aligned}$$

As for the p^f, the $p^a > 0$ condition excludes the $\tilde{p}^a_-$ point and the stationary point is:

$$\tilde{p}^a = \frac{1}{2\alpha}\left(((\alpha-1)\sigma_o^2 - \sigma_q^2) + \sqrt{((1-\alpha)\sigma_o^2 + \sigma_q^2)^2 + 4\alpha\sigma_o^2\sigma_q^2}\right) \tag{G.14}$$

For $\alpha = 1$ we obtain:

$$\begin{aligned}
\tilde{p}^a_{\alpha=1} &= \tilde{p}^f_{\alpha=1} - \sigma_q^2 \\
&= \frac{1}{2}\left(-\sigma_q^2 + \sqrt{\sigma_q^4 + 4\sigma_o^2\sigma_q^2}\right)
\end{aligned} \tag{G.15}$$

G.2.3 Stability

Given the p_k^f evolution rule:

$$p_k^f = \frac{\alpha\sigma_o^2 p_{k-1}^f}{\sigma_o^2 + p_{k-1}^f} + \sigma_q^2 \tag{G.16}$$

If we perturb the stationary point $\tilde{p}^f$ with an $\epsilon_{k-1} \ll 1$ the following value will be:

$$\begin{aligned}
\tilde{p}^f + \epsilon_k &= \frac{\alpha\sigma_o^2(\tilde{p}^f + \epsilon_{k-1})}{\sigma_o^2 + \tilde{p}^f + \epsilon_{k-1}} + \sigma_q^2 \\
&= \frac{\alpha\sigma_o^2\tilde{p}^f}{\sigma_o^2 + \tilde{p}^f + \epsilon_{k-1}} + \frac{\alpha\sigma_o^2\epsilon_{k-1}}{\sigma_o^2 + \tilde{p}^f + \epsilon_{k-1}} + \sigma_q^2 \\
&= \frac{\frac{\alpha\sigma_o^2\tilde{p}^f}{\sigma_o^2+\tilde{p}^f}}{1 + \frac{\epsilon_{k-1}}{\sigma_o^2+\tilde{p}^f}} + \frac{\frac{\alpha\sigma_o^2\epsilon_{k-1}}{\sigma_o^2+\tilde{p}^f}}{1 + \frac{\epsilon_{k-1}}{\sigma_o^2+\tilde{p}^f}} + \sigma_q^2 \\
&\simeq \frac{\alpha\sigma_o^2\tilde{p}^f}{\sigma_o^2 + \tilde{p}^f}\left(1 - \frac{\epsilon_{k-1}}{\sigma_o^2 + \tilde{p}^f}\right) + \frac{\alpha\sigma_o^2\epsilon_{k-1}}{\sigma_o^2 + \tilde{p}^f}\left(1 - \frac{\epsilon_{k-1}}{\sigma_o^2 + \tilde{p}^f}\right) + \sigma_q^2 \\
&\simeq \left(\frac{\alpha\sigma_o^2\tilde{p}^f}{\sigma_o^2 + \tilde{p}^f} + \sigma_q^2\right) + \epsilon_{k-1}\alpha\frac{\sigma_o^2}{\sigma_o^2 + \tilde{p}^f}\left(1 - \frac{\tilde{p}^f}{\sigma_o^2 + \tilde{p}^f}\right) \\
&= \tilde{p}^f + \epsilon_{k-1}\alpha\left(\frac{\sigma_o^2}{\sigma_o^2 + \tilde{p}^f}\right)^2
\end{aligned}$$

Therefore ϵ evolves as:

$$\epsilon_k = \epsilon_{k-1}\alpha\left(\frac{\sigma_o^2}{\sigma_o^2 + \tilde{p}^f}\right)^2 = \epsilon_0\left(\alpha\left(\frac{\sigma_o^2}{\sigma_o^2 + \tilde{p}^f}\right)^2\right)^k$$

The validity of the inequality $\alpha\left(\frac{\sigma_o^2}{\sigma_o^2+\tilde{p}^f(\alpha,\sigma_o^2,\sigma_q^2))}\right)^2 < 1$ for $\alpha > 0$ has been verified numerically. The left hand quantity reach his maximum value in the limit for $\frac{\sigma_o^2}{\sigma_q^2} \to \infty$ in correspondence of $\alpha = 1$. As $\frac{\sigma_o^2}{\sigma_o^2+\tilde{p}^f} < 1$, for $\alpha = 1$ the stationary point $p_{\alpha=1}^f$ is then stable $\forall\alpha > 0$

Similarly:

$$\tilde{p}^a + \epsilon_k = \tilde{p}^a + \epsilon_{k-1}\alpha\left(\frac{\sigma_o^2}{\sigma_o^2 + \sigma_q^2 + \alpha\tilde{p}^a}\right)^2$$

and also in this case $\forall\alpha > 0$ $\tilde{p}^a$ is stable.

G.2.4 $\frac{1}{\sigma_q^2}$ rescaling

We can redefine the Kalman equations dividing all the variances by σ_q^2. As we can see with such rescaling the Asymptotic behavior is fully determined by α the $\frac{\sigma_o^2}{\sigma_q^2}$ ratio:

$$\begin{aligned}\frac{\tilde{p}^f}{\sigma_q^2} &= \frac{1}{2\sigma_q^2}\left((\sigma_q^2+(\alpha-1)\sigma_o^2)+\sqrt{((1-\alpha)\sigma_o^2-\sigma_q^2)^2+4\sigma_o^2\sigma_q^2}\right)\\ &= \frac{1}{2}\left(\left(1+(\alpha-1)\frac{\sigma_o^2}{\sigma_q^2}\right)+\sqrt{\left((1-\alpha)\frac{\sigma_o^2}{\sigma_q^2}-1\right)^2+4\frac{\sigma_o^2}{\sigma_q^2}}\right)\end{aligned}$$

$$\begin{aligned}\frac{\tilde{p}^a}{\sigma_q^2} &= \frac{1}{2\alpha\sigma_q^2}\left(((\alpha-1)\sigma_o^2-\sigma_q^2)+\sqrt{((1-\alpha)\sigma_o^2+\sigma_q^2)^2+4\alpha\sigma_o^2\sigma_q^2}\right)\\ &= \frac{1}{2\alpha}\left(\left((\alpha-1)\frac{\sigma_o^2}{\sigma_q^2}-1\right)+\sqrt{\left((1-\alpha)\frac{\sigma_o^2}{\sigma_q^2}+1\right)^2+4\alpha\frac{\sigma_o^2}{\sigma_q^2}}\right)\end{aligned}$$

G.2.5 Limit $\frac{\sigma_o^2}{\sigma_q^2}\to\infty$

For $\frac{\sigma_o^2}{\sigma_q^2}\gg 1$ and $\alpha\neq 1$:

$$\begin{aligned}\frac{\tilde{p}^f}{\sigma_q^2} &= \frac{1}{2}\left(\left(1+(\alpha-1)\frac{\sigma_o^2}{\sigma_q^2}\right)+\sqrt{\left((1-\alpha)\frac{\sigma_o^2}{\sigma_q^2}-1\right)^2+4\frac{\sigma_o^2}{\sigma_q^2}}\right)\\ &= \frac{1}{2}\left(1+(\alpha-1)\frac{\sigma_o^2}{\sigma_q^2}+\sqrt{\left((1-\alpha)\frac{\sigma_o^2}{\sigma_q^2}\right)^2+1-2(1-\alpha)\frac{\sigma_o^2}{\sigma_q^2}+4\frac{\sigma_o^2}{\sigma_q^2}}\right)\\ &= \frac{1}{2}\left(1+(\alpha-1)\frac{\sigma_o^2}{\sigma_q^2}+|1-\alpha|\frac{\sigma_o^2}{\sigma_q^2}\sqrt{1+\frac{1+2(1+\alpha)\frac{\sigma_o^2}{\sigma_q^2}}{\left((1-\alpha)\frac{\sigma_o^2}{\sigma_q^2}\right)^2}}\right)\\ &\simeq \frac{1}{2}\left(1+(\alpha-1)\frac{\sigma_o^2}{\sigma_q^2}+|1-\alpha|\frac{\sigma_o^2}{\sigma_q^2}\left(1+\frac{(1+\alpha)}{(1-\alpha)^2\frac{\sigma_o^2}{\sigma_q^2}}\right)\right)\end{aligned}$$

For $\alpha>1$, $|1-\alpha|=(\alpha-1)$ and the solution is then $(\alpha-1)\frac{\sigma_o^2}{\sigma_q^2}$.
For $\alpha<1$, $|1-\alpha|=(1-\alpha)$ and the solution is a constant $\frac{1}{1-\alpha}$.

For $\alpha = 1$:

$$\frac{\tilde{p}^f_{\alpha=1}}{\sigma_q^2} = \frac{1}{2}\left(1 + \sqrt{1 + 4\frac{\sigma_o^2}{\sigma_q^2}}\right) \simeq \sqrt{\frac{\sigma_o^2}{\sigma_q^2}}$$

If we want to find the α dependence near the value $\alpha = 1$ we have to make first the limit for $\alpha \to 1$ and then the limit for $\frac{\sigma_o^2}{\sigma_q^2} \to \infty$. Redefining $\beta = \alpha - 1$ in this last case we obtain:

$$\begin{aligned}
&\frac{\tilde{p}^f_{\alpha\to 1}}{\sigma_q^2} \\
&= \frac{1}{2}\left(\left(1 + \beta\frac{\sigma_o^2}{\sigma_q^2}\right) + \sqrt{\left(\beta\frac{\sigma_o^2}{\sigma_q^2} + 1\right)^2 + 4\frac{\sigma_o^2}{\sigma_q^2}}\right) \\
&= \frac{1}{2}\left(1 + \beta\frac{\sigma_o^2}{\sigma_q^2} + \sqrt{\left(\beta\frac{\sigma_o^2}{\sigma_q^2}\right)^2 + 1 + 2\beta\frac{\sigma_o^2}{\sigma_q^2} + 4\frac{\sigma_o^2}{\sigma_q^2}}\right) \\
&= \frac{1}{2}\left(1 + \beta\frac{\sigma_o^2}{\sigma_q^2} + \sqrt{1 + 4\frac{\sigma_o^2}{\sigma_q^2} + 2\beta\frac{\sigma_o^2}{\sigma_q^2} + \left(\beta\frac{\sigma_o^2}{\sigma_q^2}\right)^2}\right) \\
&\simeq \frac{1}{2}\left(1 + \beta\frac{\sigma_o^2}{\sigma_q^2} + \sqrt{1 + 4\frac{\sigma_o^2}{\sigma_q^2}}\sqrt{1 + \frac{2\beta\frac{\sigma_o^2}{\sigma_q^2}}{1 + 4\frac{\sigma_o^2}{\sigma_q^2}}}\right) \\
&\simeq \frac{1}{2}\left(1 + \beta\frac{\sigma_o^2}{\sigma_q^2} + \sqrt{1 + 4\frac{\sigma_o^2}{\sigma_q^2}}\left(1 + \frac{\beta\frac{\sigma_o^2}{\sigma_q^2}}{1 + 4\frac{\sigma_o^2}{\sigma_q^2}}\right)\right) \\
&= \frac{1}{2}\left(1 + \beta\frac{\sigma_o^2}{\sigma_q^2} + \sqrt{1 + 4\frac{\sigma_o^2}{\sigma_q^2}} + \frac{\beta\frac{\sigma_o^2}{\sigma_q^2}}{\sqrt{1 + 4\frac{\sigma_o^2}{\sigma_q^2}}}\right) \\
&= \frac{1}{2}\left(1 + \sqrt{1 + 4\frac{\sigma_o^2}{\sigma_q^2}} + \beta\frac{\sigma_o^2}{\sigma_q^2}\left(1 + \frac{1}{\sqrt{1 + 4\frac{\sigma_o^2}{\sigma_q^2}}}\right)\right) \\
&\simeq \sqrt{\frac{\sigma_o^2}{\sigma_q^2}} + \frac{\alpha - 1}{2}\frac{\sigma_o^2}{\sigma_q^2}
\end{aligned}$$

To summarize, there are three different asymptotic behaviors for $\frac{\sigma_o^2}{\sigma_q^2} \to \infty$:

$$\frac{\tilde{p}^f}{\sigma_q^2} \simeq \begin{cases} (1-\alpha)^{-1} & \text{if } \alpha < 1 \\ \sqrt{\sigma_o^2/\sigma_q^2} + (\alpha - 1)(\sigma_o^2/\sigma_q^2)/2 & \text{if } \alpha \to 1 \\ (\alpha - 1)(\sigma_o^2/\sigma_q^2) & \text{if } \alpha > 1 \end{cases} \tag{G.17}$$

G.3 Variance Reduction

If we identify as a variance reduction the fraction $\frac{\tilde{p}^a_{\alpha=1}}{\sigma_o^2}$, we easily obtain for $\alpha = 1$, as a leading term in a $\frac{\sigma_q^2}{\sigma_o^2} \to 0$ limit:

$$\frac{\tilde{p}^a_{\alpha=1}}{\sigma_o^2} = \frac{1}{2}\left(\sqrt{\left(\frac{\sigma_q^2}{\sigma_o^2}\right)^2 + 4\frac{\sigma_q^2}{\sigma_o^2}} - \frac{\sigma_q^2}{\sigma_o^2}\right) \simeq \sqrt{\frac{\sigma_q^2}{\sigma_o^2}}$$

The limit is justified as in our data we have $\frac{\sigma_q^2}{\sigma_o^2} \approx \frac{1}{100}$, giving a variance reduction of $\approx \frac{1}{10}$ (in terms of errors the reduction will be then of $\approx \frac{1}{3}$).

G.4 A different method for the limit for $\alpha = 1$

It is easy to reconnect the both successions (p_k^f and p_k^a), for $\alpha = 1$, to the form:

$$p_{k+1} = \frac{\sigma^2 p_k}{\sigma^2 + p_k} + 1$$

where $\sigma^2 = \frac{\sigma_o^2}{\sigma_q^2}$ and $p = \left(\frac{p^a}{\sigma_q^2} + 1\right)$ or $p = \frac{p^f}{\sigma_q^2}$ respectively.

Through Wolfram's Mathematica online tool (Wolfram Alpha) it is possible to compute the general equation for k term for this succession given p_0, obtaining:

$$p_k = \frac{(\sqrt{4\sigma^2+1}p_0 + 2\sigma^2 + p_0)\left(\frac{(\sqrt{4\sigma^2+1}+1)^2}{2\sigma^4}\right)^n + (\sqrt{4\sigma^2+1}p_0 - 2\sigma^2 - p_0)\left(\frac{(\sqrt{4\sigma^2+1}-1)^2}{2\sigma^4}\right)^n}{(\sqrt{4\sigma^2+1} + 2p_0 - 1)\left(\frac{(\sqrt{4\sigma^2+1}+1)^2}{2\sigma^4}\right)^n + (\sqrt{4\sigma^2+1} + 1 - 2p_0)\left(\frac{(\sqrt{4\sigma^2+1}-1)^2}{2\sigma^4}\right)^n}$$

As $(\sqrt{4\sigma^2+1}+1)^2 > (\sqrt{4\sigma^2+1}-1)^2$, in the $k \to \infty$ limit we obtain:

$$\tilde{p}_k = \frac{(\sqrt{4\sigma^2+1}p_0 + 2\sigma^2 + p_0)}{(\sqrt{4\sigma^2+1} + 2p_0 - 1)} = \frac{1}{2}(1 + \sqrt{1+4\sigma^2})$$

This is a further confirmation of our results in G.2.1 and G.2.2.

G.5 Numerical Computation

A numerical computation of the asymptotic values so far calculated has been made.

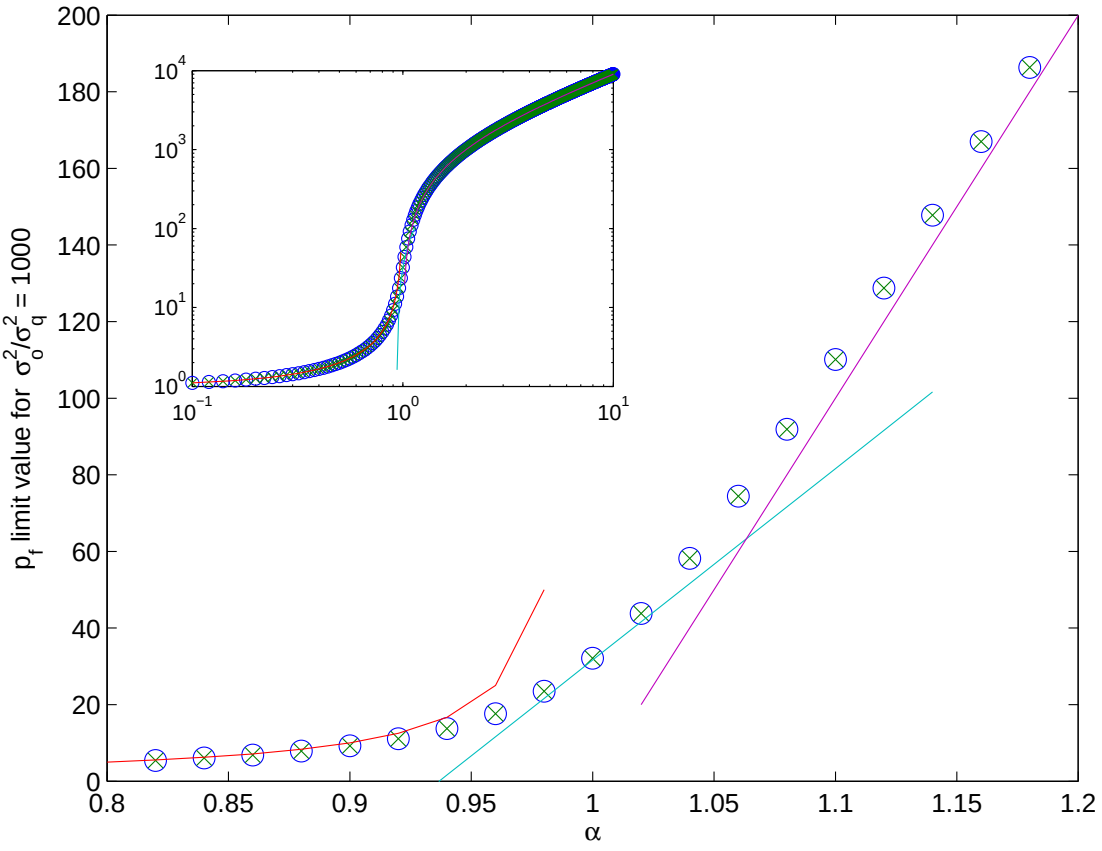


FIGURE G.1: Circles: numerical asymptotic limits. Crosses: analytical asymptotic limits. In purple: the $\alpha > 1$ approximation. In cyan: the $\alpha \to 1$ approximation. In red: the $\alpha < 1$ approximation.

G.6 Notes

- One can still calculate just the limit values for p^f and then evaluate p^a and K with the Kalman equations;
- The limit values exist also for $\alpha < 1$;
- Different behaviors are found for $\alpha < 1$, $\alpha \to 1$ and $\alpha > 1$;
- The p_- solution excluded in G.2.1 and G.2.2 is recognizable in the $\left(\frac{(\sqrt{4\sigma^2+1}-1)^2}{2\sigma^4}\right)^n$ contribution to the exact solution found in G.4;
- These results have been checked numerically.

Appendix H

Notes on 3DVAR

As sources for this appendix we refer to [75][73][76][77].

H.1 Maximum a posteriori

Given an observation vector $\mathbf{z}$ and a background (forecast) vector $\mathbf{x}^b$ which are Normally distributed with zero means and covariance matrices $\mathbf{R}$ and $\mathbf{P}^b$ respectively, the joint probability density function of the observation and background errors is given by:

$$p(\mathbf{e}^b, \mathbf{e}^r) = \frac{1}{(2\pi)^N \det(\mathbf{P}^b)^{\frac{1}{2}} \det(\mathbf{R})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{e}^b)^{\mathrm{T}}(\mathbf{P}^b)^{-1}\mathbf{e}^b - \frac{1}{2}(\mathbf{e}^r)^{\mathrm{T}}(\mathbf{R})^{-1}\mathbf{e}^r\right)$$

Where $\mathbf{e}^b = \mathbf{x}^t - \mathbf{x}^b$ and $\mathbf{e}^r = \mathbf{z} - H(\mathbf{x}^t)$.
The maximum a posteriori estimate is obtained by minimizing the cost function:

$$J(\mathbf{x}^t) = \frac{1}{2}(\mathbf{x}^t - \mathbf{x}^b)^{\mathrm{T}}(\mathbf{P}^b)^{-1}(\mathbf{x}^t - \mathbf{x}^b) + \frac{1}{2}(\mathbf{z} - H(\mathbf{x}^t))^{\mathrm{T}}(\mathbf{R})^{-1}(\mathbf{z} - H(\mathbf{x}^t))$$

H.2 Measurement & Forecast model

H.2.1 Measurement

The observation function H is supposed to be linear and the measures $\mathbf{z}$ independent, therefore:

$$\mathbf{z}_{k+1} = \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k$$

Where $\mathbf{v}_k$ is a white noise: $\langle \mathbf{v}_k \rangle = 0$ and $\langle \mathbf{v}_k \cdot (\mathbf{v}_l)^t \rangle = \mathbf{R} = \mathbf{r}\delta_l^k$ Given any linear $\mathbf{H}$ operator the cost function is purely quadratic and is guaranteed to have an unique minimum. We will assume from now on $\mathbf{H} = \mathbf{I}$. In such case the gradient of the cost functions is:

$$\nabla J(\mathbf{x}^t) = (\mathbf{P}^b)^{-1}(\mathbf{x}^t - \mathbf{x}^b) + (\mathbf{R})^{-1}(\mathbf{x}^t - \mathbf{z})$$

As the covariance matrix is diagonal the inversion of $\mathbf{R}$ is trivial.

H.2.2 Forecast

We have a time dependent linear forecast operator ϕ_k:

$$\mathbf{x}_{k+1}^b = \boldsymbol{\phi}_k \mathbf{x}_k^t + \boldsymbol{\omega}_k$$

(In the time independent case the model is stable if the eigenvalues of ϕ are less than equal to 1)
The errors $\mathbf{w}_k$ are assumed to be a white noise with $\langle \mathbf{w}_k \rangle = 0$ and
$\langle \mathbf{w}_k \cdot (\mathbf{w}_l)^t \rangle = \mathbf{q}\delta_l^k$.
At each step the forecast error covariance matrix is given by:

$$\mathbf{P}_{k+1}^b = \boldsymbol{\phi}_k \mathbf{P}_k^t \boldsymbol{\phi}_k^{\mathrm{T}} + \mathbf{q}\delta_l^k$$

The inversion of $\mathbf{P}^b$ appears non trivial.

H.3 Incremental formulation

At each step we can take the background value as a reference and calculate the increment from the background:

$$\mathbf{x}_n^t = \mathbf{x}_n^b + \delta \mathbf{x}_n$$

Thus the general 3DVAR cost function can be rewritten as:

$$\begin{aligned} J(\delta\mathbf{x}_n) &= \frac{1}{2}(\mathbf{z}-\mathbf{x}^b-\delta\mathbf{x}_n)^{\mathrm{T}}(\mathbf{R})^{-1}(\mathbf{z}-\mathbf{x}^b-\delta\mathbf{x}_n) \\ &+ \frac{1}{2}\delta\mathbf{x}_n^{\mathrm{T}}(\mathbf{P}^b)^{-1}\delta\mathbf{x}_n \\ &= J_z + J_b \end{aligned}$$

The need for $(\mathbf{P}^b)^{-1}$ can be avoided by redefining the control variable. If we can find $\mathbf{L}$ triangular such that $\mathbf{P}^b = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ then we can define a new increment $\delta\boldsymbol{\chi}_n = \mathbf{L}^{-1}\delta\mathbf{x}_n$ and minimize the cost function with respect to $\delta\boldsymbol{\chi}$:

$$J(\delta\boldsymbol{\chi}) = \frac{1}{2}(\mathbf{z}-\mathbf{x}^b-\mathbf{L}\delta\boldsymbol{\chi}_n)^{\mathrm{T}}(\mathbf{R})^{-1}(\mathbf{z}-\mathbf{x}^b-\mathbf{L}\delta\boldsymbol{\chi}_n) + \frac{1}{2}\delta\boldsymbol{\chi}_n^{\mathrm{T}}\delta\boldsymbol{\chi}_n$$

The gradient of this cost function is:

$$\begin{aligned} \nabla J(\delta\boldsymbol{\chi}) &= -\mathbf{R}^{-1}\mathbf{L}^{\mathrm{T}}(\mathbf{z}-\mathbf{x}^b-\mathbf{L}\delta\boldsymbol{\chi}_n) + \delta\boldsymbol{\chi}_n \\ &= -\mathbf{R}^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{z} + \mathbf{R}^{-1}\mathbf{L}^{\mathrm{T}}\mathbf{x}^b + (\mathbf{R}^{-1}\mathbf{P}^b + \mathbf{I})\delta\boldsymbol{\chi}_n \end{aligned}$$

H.4 Cholensky decomposition ($\mathbf{L}\mathbf{L}^{\mathrm{T}}$)

The key passage that avoids the $\mathbf{P}^b$ matrix is the decomposition in $\mathbf{P}^b = \mathbf{L}\mathbf{L}^{\mathrm{T}}$. If we write out the equation $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$

$$\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathbf{T}} = \begin{pmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{pmatrix} \begin{pmatrix} L_{11} & L_{21} & L_{31} \\ 0 & L_{22} & L_{32} \\ 0 & 0 & L_{33} \end{pmatrix}$$

we obtain the following formula for the entries of $\mathbf{L}$:

$$L_{j,j} = \sqrt{A_{j,j} - \sum_{k=1}^{j-1} L_{j,k}^2}.$$

$$L_{i,j} = \frac{1}{L_{j,j}}\left(A_{i,j} - \sum_{k=1}^{j-1} L_{i,k}L_{j,k}\right), \qquad \text{for } i > j.$$

The expression under the square root is always positive if $\mathbf{A}$ is real and positive-definite.

H.5 Optimization

H.5.1 Steepest descent

One starts with a guess $\mathbf{x}_0$ for a local minimum of F, and considers the sequence $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$ such that

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma_n \nabla F(\mathbf{x}_n),\ n \geq 0.$$

So hopefully the sequence $(\mathbf{x}_n)$ converges to the desired local minimum. Note that the value of the step size γ is allowed to change at every iteration.

H.5.2 BFGS quasi-Newton method

Newton's method and the BFGS methods need not converge unless the function has a quadratic Taylor expansion near an optimum. These methods use the first and second derivatives. In quasi-Newton methods, the Hessian matrix of second derivatives need not be evaluated directly. Instead, the Hessian matrix is approximated using rank-one updates specified by gradient evaluations (or approximate gradient evaluations). Quasi-Newton methods are a generalization of the secant method to find the root of the first derivative for multidimensional problems. In multi-dimensions the secant equation does not specify a unique solution, and quasi-Newton methods differ in how they constrain the solution. The BFGS method is one of the most popular members of this class.

From an initial guess $\mathbf{x}_0$ and an approximate Hessian matrix B_0 the following steps are repeated until x converges to the solution.

- Obtain a direction $\mathbf{p}_k$ by solving: $B_k \mathbf{p}_k = -\nabla f(\mathbf{x}_k)$.
- Perform a line search to find an acceptable stepsize α_k in the direction found in the first step, then update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k$.
- Set $\mathbf{s}_k = \alpha_k \mathbf{p}_k$.
- $\mathbf{y}_k = \nabla f(\mathbf{x}_{k+1}) - \nabla f(\mathbf{x}_k)$.
- $B_{k+1} = B_k + \frac{\mathbf{y}_k \mathbf{y}_k^{\mathrm{T}}}{\mathbf{y}_k^{\mathrm{T}} \mathbf{s}_k} - \frac{B_k \mathbf{s}_k (B_k \mathbf{s}_k)^{\mathrm{T}}}{\mathbf{s}_k^{\mathrm{T}} B_k \mathbf{s}_k}$.

$f(\mathbf{x})$ denotes the objective function to be minimized. Convergence can be checked by observing the norm of the gradient, $|\nabla f(\mathbf{x}_k)|$. Practically, B_0 can be initialized with $B_0 = I$, so that the first step will be equivalent to a gradient descent, but further steps are more and more refined by B_k, the approximation to the Hessian. The first step of the algorithm is carried out using an approximate inverse of the matrix B_k, which is usually obtained efficiently by applying the ShermanMorrison formula to the fifth line of the algorithm, giving

$$B_{k+1}^{-1} = B_k^{-1} + \frac{(\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k + \mathbf{y}_k^{\mathrm{T}} B_k^{-1}\mathbf{y}_k)(\mathbf{s}_k\mathbf{s}_k^{\mathrm{T}})}{(\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k)^2} - \frac{B_k^{-1}\mathbf{y}_k\mathbf{s}_k^{\mathrm{T}} + \mathbf{s}_k\mathbf{y}_k^{\mathrm{T}} B_k^{-1}}{\mathbf{s}_k^{\mathrm{T}}\mathbf{y}_k}$$

H.5.3 Preconditioning

All minimization algorithms work best if the iso-surfaces of the cost function are approximately spherical. The degree of sphericity of the cost function can be measured by the eigenvalues of the Hessian. (Each eigenvalue corresponds to the curvature in the direction of the corresponding eigenvector.). In particular, the convergence rate will depend on the condition number:

$$\kappa = \frac{\lambda_{max}}{\lambda_{min}}$$

If we utilize the Cholensky decomposition ($\mathbf{P}^b = \mathbf{L}\mathbf{L}^{\mathrm{T}}$) the Hessian is given by:

$$J''(\delta\boldsymbol{\chi}) = \mathbf{I} + \mathbf{L}\mathbf{R}^{-1}\mathbf{L}^{\mathrm{T}} + \ldots$$

The presence of the identity matrix in this expression guarantees that the minimum eigenvalue is ≥ 1 and there are no small eigenvalues to destroy the conditioning of the problem.

H.6 Error estimate

The optimization give us the knowledge of the most probable $\mathbf{x}_k^t$ but we need also a new estimate of the errors covariances $\mathbf{P}_k^t$.
I think we can proceed in such estimate solving the Kalman analysis equations for $\mathbf{H} = \mathbf{I}$.:

$$\begin{cases} \mathbf{x}_k^t = \mathbf{x}_k^b + \mathbf{K}_k(\mathbf{z}_k - \mathbf{x}_k^b) \\ \mathbf{P}_k^t = (\mathbf{I} - \mathbf{K}_k)\,\mathbf{P}_k^b \end{cases}$$

Obtaining:

$$\mathbf{P}_k^t = \left(\frac{\mathbf{z}-\mathbf{x}^t}{\mathbf{z}-\mathbf{x}^b}\cdot\mathbf{I}\right)\mathbf{P}_k^b$$

H.7 Further simplifications

We can assume that all the observables $\mathbf{x}$ have the same forecast errors $\omega_k$and the same measurements errors $\mathbf{v}_k$. This implies $\mathbf{r} = r$ and $\mathbf{q} = q$. Now if we divide all the covariances matrices by q obtaining $\tilde{\mathbf{P}}^t = \mathbf{P}^t/q$, $\tilde{\mathbf{P}}^b = \mathbf{P}^b/q$ and $\tilde{r} = r/q$ We can rewrite the cost functions as:

$$qJ(\mathbf{x}^t) = \frac{1}{2}(\mathbf{x}^t-\mathbf{x}^b)^{\mathrm{T}}(\tilde{\mathbf{P}}^b)^{-1}(\mathbf{x}^t-\mathbf{x}^b) + \frac{1}{2}(\mathbf{z}-\mathbf{x}^t)^{\mathrm{T}}(\tilde{r}\mathbf{I})^{-1}(\mathbf{z}-\mathbf{x}^t)$$

With $\tilde{\mathbf{P}}_k^b = \tilde{\mathbf{P}}_{k-1}^t + \mathbf{I}$ and $\tilde{\mathbf{P}}_k^t = \left(\frac{\mathbf{z}-\mathbf{x}^t}{\mathbf{z}-\mathbf{x}^b}\cdot\mathbf{I}\right)\tilde{\mathbf{P}}_k^b$.
In the minimization we can then ignore the value of q and we have just one free parameter $\tilde{r} = r/q$.
The Cholensky decomposition gives $\tilde{\mathbf{P}}^b = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^{\mathrm{T}}$ and the cost function can be incrementally rewritten as:

$$qJ(\delta\boldsymbol{\chi}) = \frac{1}{2}(\mathbf{z}-\mathbf{x}^b-\tilde{\mathbf{L}}\delta\boldsymbol{\chi}_n)^{\mathrm{T}}(\tilde{r})^{-1}(\mathbf{z}-\mathbf{x}^b-\tilde{\mathbf{L}}\delta\boldsymbol{\chi}_n) + \frac{1}{2}\delta\boldsymbol{\chi}_n^{\mathrm{T}}\delta\boldsymbol{\chi}_n \tag{H.1}$$

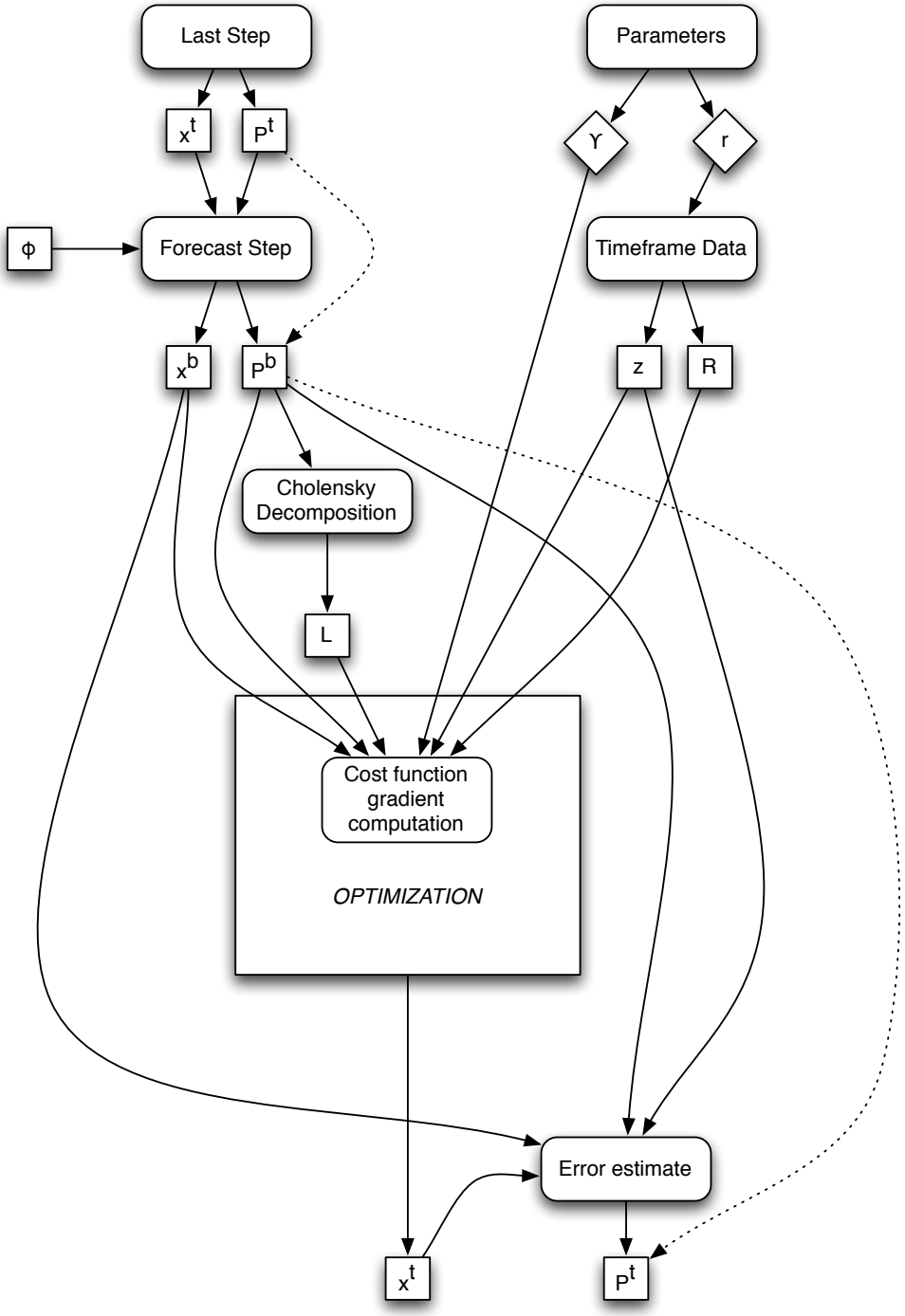
His gradient will read:

$$q\nabla J(\delta\boldsymbol{\chi}) = -\tilde{r}^{-1}\tilde{\mathbf{L}}^{\mathrm{T}}\mathbf{z} + \tilde{r}^{-1}\tilde{\mathbf{L}}^{\mathrm{T}}\mathbf{x}^b + (\tilde{r}^{-1}\tilde{\mathbf{P}}^b + \mathbf{I})\delta\boldsymbol{\chi}_n$$

This can be simplified redefining multiplying both sides for $\tilde{r}$ and explicating the value of $\mathbf{P}^b$:

$$q\tilde{r}\nabla J_k(\delta\boldsymbol{\chi}_k) = \tilde{\mathbf{L}}_k^{\mathrm{T}}\mathbf{z}_k + \tilde{\mathbf{L}}_k^{\mathrm{T}}\mathbf{x}_k^b + (\tilde{\mathbf{P}}_{k-1}^t + (\tilde{r}+1)\mathbf{I})\delta\boldsymbol{\chi}_k \tag{H.2}$$

H.8 3DVar algorithm scheme



Appendix I

Statistical Forecast

This appendix is based on [70].

I.1 Introduction

In this appendix we want to derive, from a sample of GPS data on traffic speed and fluxes over large roads, a statistical forecasting operator suitable for implementation in a Kalman Filter framework for traffic nowcasting (i.e. short term forecasting).

In fact, in the Kalman Filter plays a fundamental role the prediction of the state $\vec{x}(t+dt)$ given the vector state $\vec{x}(t)$ at time t. Using a meteorological lexicon we call here *predictors* the values of the field $\vec{x}(t)$ and *predictands* the values of $\vec{x}(t+dt)$. In our case the elements of $\vec{x}$ are the values of an observable (e.g. speed) over different points of the road network.

As a further condition, we want the forecast operator to be linear and then represented as a matrix. The values in this matrix are the coefficients of the linear combination of the predictors which form the best approximation (in the least square sense) for the predictands.

Statistical formulas have a greater probability of verifying well, when applied to new data, if the number of predictors is small relative to the number of independent observations of each predictor. To reduce the number of predictors we will project $\vec{x}$ on a base of *Empirical Orthogonal Functions* (EOF) of space $\vec{Y}$ whose coefficients

$Q(t)$ are also orthogonal functions of time. A small number of $\vec{Y}$ with large variances may then be used as predictors.

I.2 Prediction

We will develop here the statistical prediction formulas. In the most general case the predictands are different by the predictors. We will identify in this section the predictands with x and the predictors with p.

I.2.1 Single predictand

I.2.1.1 p, x: original data

Let the predictand be $x(t)$, and let the M predictors be $p_1(t), \ldots, p_M(t)$, where t is time.

For any choice of $M+1$ prediction constants $c_0, \ldots, c_M$, the prediction formula for $x(t)$ is:

$$x(t) = c_0 + \sum_{m=1}^{M} c_m p_m(t) + r(t) = \sum_{m=0}^{M} c_m p_m(t) + r(t) \tag{I.1}$$

where we have let $p_0(t) = 1$ and the final term $r(t)$, which depends upon the choice of constants, is the error in prediction $x(t)$.

The problem at hand is determining the set of $M+1$ constants c_m which minimizes the mean value of $r(t)$. This optimization is achieved taking a set of N observations of each quantity and minimizing the value of $\langle r^2 \rangle$

$$\langle r^2 \rangle = \langle x^2 \rangle - 2 \sum_{m=0}^{M} c_m \langle p_m x \rangle + \sum_{m,n=0}^{M} c_m c_n \langle p_m p_n \rangle \tag{I.2}$$

over the dataset. In order to minimize $\langle r^2 \rangle$, its derivative $\frac{d\langle r^2 \rangle}{dc_m}$ must vanish $\forall m$. Therefore we obtain as a sufficient condition:

$$\sum_{n=0}^{M} \langle p_m p_n \rangle c_n = \langle p_m x \rangle \qquad \text{for } m = 0 \ldots M \tag{I.3}$$

Solving this set of $M+1$ equations in the $M+1$ unknowns c_m we obtain the prediction constants. Equation I.3 is a necessary and sufficient condition that $\langle r^2 \rangle$ is minimized. From eq. I.1 and eq. I.3 can be easily obtained an equivalent condition:

$$\langle p_m r \rangle = 0 \qquad \text{for } m = 0 \dots M \tag{I.4}$$

I.2.1.2 $\tilde{p}, \tilde{x}$: departures from mean values

Using a prediction formula which refers to departures from their mean values of predictors $\tilde{p}$ and predictands $\tilde{x}$, is useful as we can avoid the $m = 0$ term:

$$\tilde{x}(t) = \sum_{m=1}^{M} c_m \tilde{p}_m(t) + r(t) \tag{I.5}$$

and may permit us to easily quantify the quality of the predictor through a useful quantity called *reduction of variance*.

The equations for the prediction constants in eq. I.5 are then:

$$\sum_{n=1}^{M} \langle \tilde{p}_m \tilde{p}_n \rangle c_n = \langle \tilde{p}_m \tilde{x} \rangle \qquad \text{for } m = 1 \dots M \tag{I.6}$$

where $\langle \tilde{p}_m \tilde{p}_n \rangle$ and $\langle \tilde{p}_m \tilde{x} \rangle$ are the sample covariances with respect to time. Also in this case stands the equivalent condition:

$$\langle \tilde{p}_m r \rangle = 0 \qquad \text{for } m = 1 \dots M \tag{I.7}$$

Moreover, from eq. I.5 and eq. I.7 follows that:

$$\langle r^2 \rangle = \langle \tilde{x}^2 \rangle - \langle \sum_{n=1}^{M} c_n \tilde{p}_m \rangle \tag{I.8}$$

The last formula describes how the unexplained variance (on the left) is equal to the variance of the predictand x minus the amount of the variance of x explained by the predictors. The ratio

$$rv = \frac{\langle \sum_{n=1}^{M} c_n \tilde{p}_m \rangle}{\langle \tilde{x}^2 \rangle} \tag{I.9}$$

is therefore called *reduction of variance* and is used, as we said, as a measure of the goodness of the prediction.

I.2.1.3 $\hat{p}, \hat{x}$: standardized signals

If we standardize predictand and predictors we obtain a system of conditions involving correlations instead of covariances:

$$\sum_{n=1}^{M} \langle \hat{p}_m \hat{p}_n \rangle c_n = \langle \hat{p}_m \hat{x} \rangle \qquad \text{for } m = 1 \dots M \tag{I.10}$$

and the reduction of variance is directly:

$$rv = \langle \sum_{n=1}^{M} c_n \hat{p}_m \rangle \tag{I.11}$$

I.2.2 Sampling and over-fitting issues

The means $\langle p_m p_n \rangle$ and $\langle p_m x \rangle$, as the covariances $\langle \tilde{p}_m \tilde{p}_n \rangle$ and $\langle \tilde{p}_m \tilde{x} \rangle$ and correlations $\langle \hat{p}_m \hat{p}_n \rangle$ and $\langle \hat{p}_m \hat{x} \rangle$ tend to differ considerably from one sample to another, and hence from sample to population. It follows that the coefficients c_m might also depend upon the particular sample. Therefore, any effort to ensure that a prediction formula is the best for a particular sample, rather than merely good, are probably wasted. Indeed, a formula which appears good for one sample may be poor for the population. At first glance it might seem that a greater number of predictors should lead to a greater probability of obtaining a good prediction formula. This would be so if the sample used in establishing the formula was consistent with the entire population. But the greater the number of predictors, the greater the probability that some linear combination of these predictors will be highly correlated with the predictand within the sample, even though it may be uncorrelated with the predictand within the population.

If we assume that population means do exist, we may let S_0 and R_0 be the reduction of variance and the ratio of unexplained variance to the total variance within the population ($S_0 + R_0 = 1$). We may also let S' be the expected reduction of variance within the sample where the prediction formula is the best, and let S'' be the expected reduction of error when the formula is applied to another sample. It

can be shown that:

$$S' = S_0 + \frac{M}{N-1} R_0 \tag{I.12}$$

$$S'' = S_0 - \frac{M}{N+1} R_0 \tag{I.13}$$

Thus a considerable discrepancy is expected between the reduction of variance and the reduction of error and this discrepancy is proportional to the ratio M/N. Therefore to ensure we have picked an efficient formula the sample should be as great as possible and the number of predictors restricted. Having a great number of predictors gives the maximum information, but makes higher the danger of sampling problems.

I.2.3 Multiple predictands formula

If we have a vector of D predictands $\vec{x}$, the formula *I.6* is valid $\forall x_i$ and can be rewritten in matrix sense (where repetition of indices implies summation) as:

$$\langle \tilde{p}_m \tilde{p}_n \rangle C_{n,i} = \langle \tilde{p}_m \tilde{x}_i \rangle \qquad \text{for } m, n = 0 \dots M \text{ and } i = 0 \dots D \tag{I.14}$$

Furthermore, being this a set of D linear system, it is possible to evaluate the matrix coefficients $C_{n,i}$ through the inversion of the square matrix $\langle \tilde{p}_m \tilde{p}_n \rangle$:

$$C_{n,i} = \langle \tilde{p}_m \tilde{p}_n \rangle^{-1} \langle \tilde{p}_m \tilde{x}_i \rangle \qquad \text{for } m, n = 0 \dots M \text{ and } i = 0 \dots D \tag{I.15}$$

I.3 Dimensionality Reduction

In I.2.2 it has been seen how the difference between the reduction of variance in the sample and the reduction of variance in the population proportional to the ratio M/N, being M the number of predictors and N the sample size. It is therefore important to reduce the number of predictors for our statistical forecast formula. First, we can simply avoid keeping in our analysis field which are poor of information or considered less important (in our traffic application these could be secondary roads with only a few data per hour). When the fundamental core of our predictors has been identified, it is then possible to make a further reduction of the dimensionality of the problem, retaining in the meanwhile the dimensionality

of the problem, describing in an approximate way the fields in analysis through a set of *Empirical Orthogonal Functions*.

I.3.1 Empirical Orthogonal Functions

We want to determine a set of quantities, in number smaller than the number of given predictors, such that all predictors may be then be approximated by linear combinations of these new quantities. Then we can reduce the number of predictors, using as predictors these new quantities. To find these quantities (called in meteorology Empirical Orthogonal Functions and in other field Principal Components) we consider a set of M predictors $p_1(t), \ldots, p_M(t)$, each observed N times $t_1, \ldots, t_n$. We call *total variance* of the predictors the sum:

$$V = \sum_{m=1}^{M} \langle p_m^2 \rangle \tag{I.16}$$

Let $q_1(t), \ldots, q_K(t)$ be any K quantities, where $K < M$ and let

$$p_m^*(t_i) = \sum_{k=1}^{K} y_{k,m} q_k(t_i) + r_m(t_i) \tag{I.17}$$

Where the $y_{k,m}$ have to be chosen to minimize the *unexplained variance*:

$$R = \sum_{m=1}^{M} \langle r_m^2 \rangle \tag{I.18}$$

The value of R remains function of the choice of the q_k. The problem at hand is then choosing the right q_k to minimize R, and thus maximize the quantity $(V-R)/V$, that becomes the fraction of total variance which may be represented by K quantities. We proceed in this choice projecting the p_k on a base of orthogonal functions:

$$p_m(t) = \sum_{k=1}^{M} Y_{k,m} Q_k(t) \tag{I.19}$$

Where the $Y_{k,m}$ are orthonormal functions of space:

$$\sum_{k=1}^{M} Y_{k,m} Y_{j,m} = \delta_{k,j} \tag{I.20}$$

and the $\tilde{Q}_k(t)$ (where the tilde means again the fluctuations from the mean value in time) are orthogonal functions of time:

$$\langle \tilde{Q}_k \tilde{Q}_j \rangle = a_k \delta_{k,j} \tag{I.21}$$

with $a_k \geq a_{k+1} \geq 0$. It is proven [70] that the quantities $\tilde{Q}_1, \ldots, \tilde{Q}_K$ minimize the error R. In this case the variance can be easily estimated with the following formulas:

$$V = \sum_{k=1}^{M} a_k \tag{I.22}$$

$$R = \sum_{k=K+1}^{M} a_k \tag{I.23}$$

$$V - R = \sum_{k=1}^{K} a_k \tag{I.24}$$

In order to describe a method for determining the quantities $Y_{k,m}$ and $Q_k(t)$ satisfying (I.19) it is convenient to use matrix notation. Let then P,$\tilde{P}$,Q,$\tilde{Q}$ be matrices of N rows and M columns whose elements are respectively $p_m(t_i)$,$\tilde{p}_m(t_i)$,$Q_k(t_i)$,$\tilde{Q}_k(t_i)$, and let Y be a square matrix of order M whose elements are $Y_{k,m}$. The problem consists in expressing P in the form:

$$\begin{cases} P &= QY \\ Y^t Y &= I \\ \tilde{Q}^t \tilde{Q} &= D \end{cases} \tag{I.25}$$

where I is the identity matrix and D a diagonal matrix whose diagonal elements are the decreasing a_k. This system can be rewritten as:

$$\begin{cases} Q &= PY^t \\ Y^t Y &= I \\ Y\tilde{P}^t \tilde{P} Y^t &= D \end{cases} \tag{I.26}$$

Therefore, if Y satisfies the last two equations in (I.26) then Q is defined by the first. Let:

$$A = \tilde{P}^t \tilde{P} \tag{I.27}$$

be a matrix whose elements $N\langle p_j^* p_k^* \rangle$ are proportional to the time covariances of the predictors. Thus the problem of finding Y becomes a simple eigenvalue problem

of the covariance matrix A:

$$YAY^t = D \tag{I.28}$$

while, as just said, Q can be than determined from the equation:

$$Q = PY^t \tag{I.29}$$

I.3.2 Role of Noise

Each of the p_k is likely to contain some noise. The Q_k with small variances may be regarded as the small residuals in approximate linear relations connecting the p_k, therefore they are likely to consist almost entirely of noise, like many other quantities which are small differences between larger quantities. These same remarks do not apply to the Q_k with large variances, since, although they may contain as much noise as the other Q_k, they should contain less noise relative their total variance.

I.3.3 Alternative formulations

We have seen in section I.2.1 that, in developing a prediction formula, signals can be taken in three different ways: original data, departures from mean values and standardized signals. Depending on which signal we want to use as a predictor, the problem can be redefined.

For instance if we want to use original data instead of the covariance matrix (multiplied by N), we should let $A = P^tP$. In this case has been observed that $\sum_{k=1}^{M} a_k$ is not anymore the explained variance, while $Y_{1,m}$ represents $\langle p_m \rangle$. The advantage of this approach is only the straightforward procedure, as it does not involve the addition or subtraction of the mean values (or the rescaling of the fluctuations).

For standardized signals instead we have $A = \hat{P}^t\hat{P}$, i.e. the correlation matrix, multiplied again by N. In this kind of approach we increment the relative weights of low variance signals, and therefore the possibility of a more accurate prediction for these fields. On the other hand, with a more sensible information extraction for those signals we are also amplifying noise. Thus this approach should be considered only for systems where little fluctuations in low variance signals are of the same relevance as considerable variations high variance ones.

In general, the covariance case we have illustrated so far is often the most suitable for a nonspecific problem. Furthermore, in our specific case, if we consider the speed fields over a road network, their mean values are reasonably comparable. Therefore fluctuation in different roads can be compared without the need of a rescaling and is not worth to try to focus our forecast over signals where fluctuations are negligible. For this reason, from now on, we will always use as signals $\tilde{P}$ and as time EOF $\tilde{Q}$. When finally we will want to describe the original data will sufficient to add to each these functions the mean value of the signal.

I.3.4 Space-time inversion

Besides the space base Y, we want also the time coefficient $\tilde{Q}$ to be orthonormal basis of time. We should then redefine it as:

$$\tilde{Q} = \tilde{Q} D^{\frac{1}{2}} \tag{I.30}$$

and with this transformation EOF system becomes then equivalent to:

$$\begin{cases} \tilde{P} &= \tilde{Q} D^{\frac{1}{2}} Y \\ Y^t Y &= I \\ \tilde{Q}^t \tilde{Q} &= I \end{cases} \tag{I.31}$$

And the solution can be now found with:

$$\begin{cases} \tilde{Q} &= \tilde{P} Y^t D^{-\frac{1}{2}} \\ Y^t Y &= I \\ D^{-\frac{1}{2}} Y \tilde{P}^t \tilde{P} Y^t D^{-\frac{1}{2}} &= I \end{cases}$$

$$\begin{cases} \tilde{Q} &= \tilde{P} Y^t D^{-\frac{1}{2}} \\ Y^t Y &= I \\ Y A Y^t &= D \end{cases} \tag{I.32}$$

We will now show how it is possible to invert the problem and find firstly the $\tilde{Q}$ as EOF of time, then the Y as their coefficients in space. The way for doing this is to transpose the whole problem:

$$\begin{cases} \tilde{P}^t &= Y^t D^{\frac{1}{2}} \tilde{Q}^t \\ Y Y^t &= I \\ \tilde{Q} \tilde{Q}^t &= I \end{cases} \tag{I.33}$$

$$\begin{cases} Y^t &= \tilde{P}^t \tilde{Q} D^{-\frac{1}{2}} \\ D^{-\frac{1}{2}} Q^t \tilde{P} \tilde{P}^t Q D^{-\frac{1}{2}} &= I \\ \tilde{Q}\tilde{Q}^t &= I \end{cases}$$

$$\begin{cases} Y^t &= \tilde{P}^t Q D^{-\frac{1}{2}} \\ Q^t B^t Q &= D \\ \tilde{Q}\tilde{Q}^t &= I \end{cases} \tag{I.34}$$

Here $B = \tilde{P}\tilde{P}^t$ is a matrix whose elements are proportional to the space covariances of the signals(and not time covariances as A) : $B_{j,k} = \sum_{m=1}^{M} \tilde{p}_m(t_i)\tilde{p}_m(t_k)$. We can notice that the equations in (I.32) have the same shape of (I.34). In particular, the second equation:

$$Q^t B^t Q = D \tag{I.35}$$

have the shape of an eigenvalue problem. But, if we check the dimensions of the matrices involved, we have:

$$(M \times N)(N \times N)(N \times M) = (M \times M) \tag{I.36}$$

where M is the space dimension and N the time dimension. We have then 3 possible situations, which have been analyzed numerically:

- for $N = M$ the eigenvalue problem is again well defined and we can obtain the same space and time EOF both diagonalizing the A or B;
- for $N < M$ we can find at maximum N independent eigenvectors of B, associated to N eigenvalues in D, and the others $M - N$ values of the diagonal of D are 0;
- for $N > M$ we find N independent eigenvectors of B, where M are associated to the M non-zero eigenvalues in D and $N - M$ are associated to the eigenvalue 0.

In conclusion, it has been shown that, from a data matrix P of dimension $N \times M$, only a number of EOF equal to the minimum value between M and N can be extracted. The space and time EOF can be then computed both solving the $A = \tilde{P}^t\tilde{P}$ $(M \times M)$ and the $B = \tilde{P}\tilde{P}^t$ $(N \times N)$ eigenvalue problem. Therefore, if N and M have a significant difference, the choice of which matrix have to be memorized and then diagonalized.

I.4 Final Formulas

We want here to obtain a final version of the prediction formula, suitable for a direct application to our data. Our aim is then to make the formula operating directly to a field (in our case the speed field $v(t)$), in order to forecast the state after a time dt. Let $x = v(t + dt)$ be the predictands and $p = v(t)$ the predictors. Let also be $\tilde{x} = x - \langle v \rangle$ and $\tilde{p} = p - \langle v \rangle$. Predictands and predictors have been projected to a truncated basis of $K < M$ EOF:

$$x = q^x Y \tag{I.37}$$

$$\tilde{x} = x - \langle v \rangle = \tilde{q}^x Y = (q^x - \langle q \rangle) Y = q^x Y - \langle q \rangle Y \tag{I.38}$$

$$\langle q \rangle = \langle q^x \rangle = \langle q^p \rangle = \langle v \rangle Y^t \tag{I.39}$$

And the forecast matrix C has to be computed over the coefficients q:

$$C_{n,i} = \langle \tilde{q}^p_m \tilde{q}^p_n \rangle^{-1} \langle \tilde{q}^p_m \tilde{q}^x_i \rangle \qquad \text{for } m, n = 0 \dots K \text{ and } i = 0 \dots D \tag{I.40}$$

where D is the dimension of the sample where the predictor is created. In the formulation so far developed we have let the predictand x, the predictor vector p be row vectors of dimensions $1 \times M$, their coefficients q^x and q^p be row vectors of dimensions $1 \times K$, while C is a $K \times K$ matrix and Y a $K \times M$ matrix. Then, the full forecast formula has to be obtained through 3 passages, representing the transformation of operator C in the EOF space:

Projecting: $\tilde{q}^p = \tilde{p} Y^t$

Forecasting: $\tilde{q}^x = \tilde{q}^p C = \tilde{p} Y^t C$

Reconstruction: $\tilde{x} = \tilde{q}^x Y = \tilde{p} Y^t C Y$

Finally, we can rewrite the equations for data arranged in column vectors. If we let now be $v(t + dt) = x^t$, $v(t) = p^t$ and $\langle v \rangle$ be column vectors of M elements. After transposing all the equation we obtain as forecast formula:

$$\tilde{v}(t + dt) = \phi \tilde{v}(t) \tag{I.41}$$

$$\tag{I.42}$$

where the forecast matrix is:

$$\phi = Y^t C^t Y \tag{I.43}$$

Bibliography

[1] M. Barthélemy, *Spatial networks*, Physics Reports **499**, Issues 1-3, Pages 1-101, (2011).

[2] D. Balcan, V. Colizza, B. Gonçalves, H. Hud, J.J. Ramasco and A. Vespignani, *Multiscale mobility networks and the spatial spreading of infectious diseases*, Proc. Nat. Acad. Sci. **106**, no. 51, 21484, (2009).

[3] D. Helbing, *Traffic and related self-driven many-particle systems*, Rev. Mod. Phys. **73**, pp. 1067-1141, (2001).

[4] P. Ball, *Why Society is a Complex Matter*, Springer, Berlin, (2012).

[5] J. Carcopino, *Daily life in ancient Rome: the people and the city at the height of the empire*, Yale University Press, New Haven (Conn.), (1940).

[6] D. Helbing, S. Bishop, P. Lukowicz and the FuturICT Consortium, *FuturICT*, arXiv:1211.2313, (2012).

[7] A. Vespignani, *Modelling dynamical processes in complex socio-technical systems*, Nature Physics **8**, 32, (2012).

[8] D. Helbing, *Social Self-Organization*, Springer, Berlin, (2012).

[9] A. Vespignani, *Predicting the Behavior of Techno-Social Systems*, Science **325**, 425-428, (2009)

[10] http://traffico.octotelematics.it

[11] I. Benenson, K. Martens and S. Birfir, *PARKAGENT: An agent-based model of parking in the city*, Computers, Environment and Urban Systems **32**, 431-439, (2008).

[12] L. Giovannini, *A Novel Map-Matching Procedure for Low-Sampling GPS Data with Applications to Traffic Flow Analysis*, PhD Thesis, Dept. Physics, University of Bologna, (2010).

[13] D. Brockmann, L. Hufnagel and T. Geisel, *The scaling laws of human travel*, Nature **439**, 462-465, (2006).

[14] M.C. González, C.A. Hidalgo and A.L. Barabási, *Understanding individual human mobility patterns*, Nature, **453**, 779-782, (2008).

[15] A. Bazzani, B. Giorgini, R. Gallotti, L. Giovannini and S. Rambaldi, *Statistical laws in urban mobility from microscopic GPS data in the area of Florence*, J. Stat. Mech. P05001, (2010).

[16] A. Poekt, J.R. Banavar, A. Maritan and D.W. Pfaff, *Scale invariance in the dynamics of spontaneous behavior*, Proc. Natl. Acad. Sci. **109**, 10564, (2012).

[17] C. Song, T. Koren, P. Wang and A.L. Barabási, *Modelling the scaling properties of human mobility*, Nature Physics, (2010).

[18] R. Gallotti, A. Bazzani and S. Rambaldi, *Toward a Statistical Physics of Human Mobility*, IJMPC **23**, No 09, (2012).

[19] L. Pietronero, E. Tosatti, V. Tosatti and A. Vespignani, *Explaining the uneven distribution of numbers in nature: The Laws of Benford and Zipf*, Physica A: Statistical Mechanics and its Applications, **293** (12), pp. 297-304, (2001).

[20] A. Proekt, J.R. Banavar, A. Maritan and D.W. Pfaff, *Scale invariance in the dynamics of spontaneous behavior*, Proc. Nat. Acad. Sci. **109**, 10564, (2012).

[21] S. Schönfelder and K.W. Axhausen, *Structure and innovation of human activity space*, Arbeitsbericht Verkerhrs und Raumplanung **258**, IVT ETH Zurich, (2004).

[22] P. S. Samuelson and W. D. Nordhaus, *Economics 17th Edition*, McGraw-Hill, (2004).

[23] A.L. Barabási, *The origin of bursts and heavy tails in human dynamics*, Nature **435**, 207-211, (2005).

[24] A. Vázquez, J. Gama Oliveira, Z. Dezsö, K-I. Goh,I. Kondor and A.L. Barabási, *Modeling bursts and heavy tails in human dynamics*, Physical Review E **73**, 036127, (2006).

[25] D.B Stouffer, R.D. Malmgren, and L.A.N Amaral, *Comment on "The origin of bursts and heavy tails in human dynamics"*, arXiv:physics/0510216, (2005).

[26] A.L. Barabási, K.I. Goh, A. Vázquez, *Reply to Comment on "The origin of bursts and heavy tails in human dynamics"*, arXiv:physics/0511186, (2005)

[27] N. Eagle, A. Pentland, and D. Lazer, *Inferring Social Network Structure using Mobile Phone Data*, PNAS **106** (36), pp. 15274-15278 (2009).

[28] http://www.imggot.com/

[29] S. Dehaene, *The neural basis of the Weber-Fechner law: A logarithmic mental number line.*, Trends in Cognitive Sciences **7**, 145, (2003).

[30] S.S. Stevens, *On the Psychophysical Law*, Psychological Review **64** (3), 153, (1957).

[31] H. Eisler, *Experiments on subjective duration 1868-1975: A collection of power function exponents*, Psychological Bulletin **83**, 1154-1171, (1976).

[32] P. Fraisse, *Perception and estimation of time*, Annu. Rev. Psychol. **35**, 1-37, (1984).

[33] T. Takahashi, H. Oonob and M.H.B. Radfordb, *Psychophysics of time perception and intertemporal choice models*, Physica A **387**(8-9), 2066-2074, (2007).

[34] T. Takahashi, *Loss of self-control in intertemporal choice may be attributable to logarithmic time-perception*, Medical hypotheses **65**, 691, (2005).

[35] G. Zauberman, B.K. Kim, S.A. Malkoc and J.R. Bettman, *Discounting Time and Time Discounting: Subjective Time Perception and Intertemporal Preferences*, Journal of Marketing Research: Vol. **46**, No. 4, pp. 543-556, (2009).

[36] F. Simini, M.C. González, A. Maritan and A.L. Barabási, *A universal model for mobility and migration patterns*, Nature **484**, 96, (2012).

[37] R. Kölbl and D. Helbing, *Energy laws in human travel behaviour*, New J. Phys. **5**, 48.1-48.12, (2003).

[38] L.D. Landau, E.M. Lifshitz, *Statistical Physics - Course of Theoretical Physics*, Butterworth-Heinemann, Third edition, (1980).

[39] Y. Zahavi and A. Talvitie, *Regularities in Travel Time and Money Expenditures*, Transportation Research Record **750**, (1980).

[40] Y. Zahavi, *Traveltime Budget and Mobility in Urban Areas*, Washington DC, USA: US Department of Transportation, (1974).

[41] P.L. Mokhtarian and C Chen, *TTB or not TTB, that is the question: a review and analysis of the empirical literature on travel time (and money) budgets*, Transport Research A, **38** (9-10), pp. 643-675 ,(2004).

[42] B. van Wee, P. Rietveld and H. Meurs, *Is average daily travel time expenditure constant? In search of explanations for an increase in average travel time*, Journal of Transport Geography, **14**(2), 109-122, (2006).

[43] http://sitis.istat.it/sitis/html/

[44] http://www.ilsole24ore.com/speciali/redditi_comuni_08/index.shtml

[45] http://www.immobiliare.it/

[46] A.L. Barabási and R. Albert, *Emergence of scaling in random networks*, Science **286**, 509, (1999).

[47] R. Gallotti, *Mobilità Urbana: individui razionali su una rete complessa*, MSc Thesis, University of Bologna, (2009).

[48] G.K. Zipf, *Human behaviour and the principle of least effort*, Addison-Wesley, Boston, (1949).

[49] R. Baeza-Yates and G. Navarro, *Block addressing indices for approximate text retrieval*, Journal of the American Society for Information Science **51**, 69, (2000).

[50] H.S. Heaps, *Information Retrieval: Computational and Theoretical Aspects*, Academic Press, Orlando (1978).

[51] C. Song, Z. Qu, N. Blumm and A.L. Barabási, *Limits of Predictability in Human Mobility*, Science, **327**, 1018-1021, (2010).

[52] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki and C. Ratti, *Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data*, Inter. Conf. on Pattern Recognition (ICPR 2010), Workshop on Human Behavior Understanding (HBU), pp. 1425. Springer, Berlin, (2010).

[53] J.P. Bagrow, Y.-R. Lin, *Spatiotemporal features of human mobility*, arXiv:1202.0224, (2012)

[54] A. Noulas, S. Scellato, R. Lambiotte, M. Pontil and C. Mascolo, *A tale of many cities: universal patterns in human urban mobility*, arXiv:1108.5355v4, (2011)

[55] D. Helbing, *Quantitative Sociodynamics*, Springer, Berlin, 18-23, (2010).

[56] D. Helbing, *Verkehrsdynamik. Neue physikalische Modellierungskonzepte*, Springer, Berlin,143-145, (1997).

[57] S. Rambaldi, A. Bazzani, B. Giorgini and L. Giovannini, *Mobility in modern cities: looking for physical laws*, Proc. ECCS07 Conf., paper n. 132, (2007).

[58] I. Illich, *Energy and Equity*, The Trinity Press, London, (1974)

[59] K. Gyimesi, C. Vincent, and N. Lamba, *Frustration Rising*, IBM Global Commuter Pain Survey, (2011).

[60] D. McFadden, *Conditional logit analysis of qualitative choice behavior*, in Frontiers in Econometrics, ed. by P. Zarembka, New York: Academic Press, 105-142, (1974)

[61] J.C. Wardrop, *Some Theoretical Aspects of Road Traffic Research*, Proceedings Institution of Civil Engineers, London, Part **2, 9**, 325, (1952).

[62] A. Haurie and P. Marcotte, *On the Relationship between Nash-Cournot and Wardrop Equilibria*, Networks **15**, 295-308, (1985).

[63] D. Monderer and L.S. Shapley, *Potential Games*, Games and Economic Behavior **14**, 124-143, (1996).

[64] T. Roughgarden and É. Tardos, *How Bad is Selfish Routing?*, Journal of the ACM **49** No 2, 236-259 ,(2002).

[65] J.D. Murchland, *Braess's Paroadox of Traffic Flow*, Traspn. Res., Vol **4**, pp 391-394, (1970).

[66] E. Manley, T. Cheng and A. Emmonds, *Understanding route choice by using agent-based simulation*, Proceedings of the 11th international conference on GeoComputation. (pp. 54 - 58). University College London, (2011).

[67] http://pegasus.octotelematics.com

[68] S. Rambaldi, M. Marchioni, A. Bazzani and B. Giorgini area of Rome, *Traffic Global Analysis on the Whole Italian Road Network*, MIPRO, 2012 Proceedings of the 35th International Convention, (2012).

[69] D.B. Work S. Blandin, O-P. Tossavainen, B. Piccoli, and A. M. Bayen, *A Traffic Model for Velocity Data Assimilation*, Applied Mathematics Research eXpress, Vol. **2010**, No. 1, pp. 135, (2010)

[70] E. Lorenz, *Empirical orthogonal functions and statistical weather prediction*, Tech. Rep. **1**, Statistical Forecasting Project, Dept. Meteorology, MIT, 49 pp., (1956)

[71] C. Poletto, M. Tizzoni and V. Colizza, *Heterogeneous length of stay of hosts' movements and spatial epidemic spread*, Scientific Reports **2**:476, (2012).

[72] I. Volkov, J.R. Banavar, S.P. Hubbell and A. Maritan, *Inferring species interactions in tropical forests*, Proc. Nat. Acad. Sci. **106**, 13854, (2009).

[73] S. Polavarapu, *Atmospheric Data Assimilation*, Lecture notes.

[74] M. Pilolli, *A Dynamical System Approach to Data Assimilation in Chaotic Models*, PhD Thesis, Dept. Physics, University of Bologna, (2008).

[75] P. Courtier, J.-N. Pawand and A. Hollingsworth, *A strategy for operational implementation of 4D-Var, using an incremental approach*, Q.R.Meteorol.SOC., 120,pp.1367-1387, (1994)

[76] M. Fisher, *Lecture notes on Data Assimilation*, European Center for Medium Range Weather Forecast, (2010).

[77] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*, Cambridge University Press New York, NY, USA, (2007).