UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI FISICA E ASTRONOMIA
Corso di Laurea Magistrale in Fisica

Tesi

Validation of Statistical Model of Spatial Flows

Relatore
Prof. Amos Maritan

Correlatori
Prof. Filippo Simini, Bristol University
Prof. Samir Suweis

Candidata
Irene Malvestio

Anno Accademico 2014-2015

Contents

Introduction

Understanding and modeling the general patterns of human mobility is a long-standing problem in sociology and human geography. It is not only a major intellectual challenge, but also of importance for public health, city planning, traffic engineering and economic forecasting. For example, quantifiable models of human mobility are indispensable for predicting the spread of biological pathogens.

Research in this area gained new perspectives, arousing the interest of physicists ([BHG06], [GHB08]) due to the availability of several accurate and large scale electronic dataset, which helps tracking the mobility fluxes and thus allowing to test the the hypotheses and results of different models.

Statistical models of spatial flows have been traditionally developed starting from the principle of entropy maximization subject to various constraints such as the presence of a finite amount of resources to move around. Traditionally mobility fluxes were described by models originating from physics. The mostly widely is the *Gravity Model* ([Zip46]) that postulates fluxes in analogy with the Newtons law of gravitation, where the number of commuters between two locations is proportional to their populations (i.e. the demographic mass) and decays with the square of the distance between them.

Besides the Gravity Model, several other models were used like the *Intervening Opportunities model* ([Sto40]) or the parameter-free *Radiation model* ([SGMB12]). Despite the new alternatives, the *Gravity Model* (with variations and improvements) is still the prevailing framework used to predict population movement [Bar10] [TTG+10] [JWS08], cargo shipping volume [KKGB10] and inter-city phone calls [KCRB09], [EEBL].

All the most commonly used approaches result in estimating the flows as the product of two types of variables, one type that depends on an attiribute of each single location (e.g. the population), and the other type that depends on a quantity relating a pair of locations (e.g. the distance or travel time). The differences between the various models consist of the choice of variables considered and the specific function of the distance. The aim of this research is to develop a set of statistical methods based on non-parametric regression and scaling techniques that will allow us to understand whether a given class of models is compatible with a set of observed flows, irrespective of the particular choice of the variables and functional forms. The proposed methodology is tested on synthetically generated data, and applied to describe empirical data of commuting and relocation flows in the United States, and commuting flows in England.

In the first chapter we present studies of individual mobility based on data of dispersal of bank notes and cell phone records, modeled with continuous time random walks. In the

second chapter we show some formulations of the gravity model, we derive the models by maximising the entropy, and we describe two alternatives: the *Intervening Opportunities* and the *Radiation Model*. In the third chapter we present the *Generalized Linear Model*, which is a technique for fitting the Gravity Model, and we apply it to commuting flows in the USA. In the fourth chapter we show the methods that we use for non-parametric fit, and we test them on synthetic data. In the fifth chapter we present some techniques for testing spatial model assumptions that we apply in the following chapter.

Chapter 1

Individual human mobility patterns

Given the many unknown factors that influence population mobility patterns, ranging from means of transportation to job- and family-imposed restrictions and priorities, human trajectories are often approximated with various random walk or diffusion models. In this chapter we will describe some approaches that physicists have developed to describe human mobility, studying data of dispersal of bank notes or mobile phone records.

1.1 Dispersal of bank notes

Brockmann et al. ([BHG06]) analysed the dispersal of bank notes to conclude that travelling behaviour can be described by a continuous time random walk process incorporating scale free jumps as well as long waiting times between displacements.

They utilise data collected at online bill tracking websites which monitor the worldwide dispersal of large numbers of individual bank notes and then infer the statistical properties of human dispersal with very high spatiotemporal precision. Their analysis of human movement is based on trajectories of $464,670$ dollar bills obtained from the bill tracking system www.wheresgeorge.com. They analysed the dispersal of bank notes in the United States, excluding Alaska and Hawaii. The core data consists of 1,033,095 reports to the bill tracking website. From these reports they calculated the geographical displacements $r = |x_2 - x_1|$ between a primary (x_1) and secondary (x_2) report location of a bank note and the elapsed time T between successive reports.

From a total of 20,540 short time trajectories originating all across the United States they measured the probability $p(r)$ of traversing a distance r in a time interval δT between one and four days. Between a radius $L_{min} = 10km$ and the approximate average East-West extension of the United States $L_{max} \sim 3200km$, the distribution exhibits power law behaviour $p(r) \sim r^{-(1+\beta)}$ with an exponent $\beta = 0.59 \pm 0.02$. For $r < L_{min}$, $p(r)$ increases linearly with r which implies that displacements are distributed uniformly inside the disk. They measured $p(r)$ for three classes of initial entry locations, big, intermediate, and small cities, and they found that all distributions exhibit an algebraic tail with the same exponent $\beta \sim 0.6$. If the trajectories are considered random walks, an exponent β in the interval $0 < \beta < 2$ is characteristic of Lévy flights (see next section for more details). With this assumption, it is possible to estimate the time T_{eq} for an initially localised ensemble of bank notes to reach

the stationary distribution. In this case they obtained $T_{eq} \sim 68$days, but data show a far lesser dispersion than expected. To investigate this problem, they considered the relative proportion $P_0^i(t)$ of bank notes which are reported in a small (20 km) radius around the initial entry location i as a function of time, and used this quantity as an estimate of the probability of a bank note being reported at the initial location at time t (first-passage times). They found the asymptotic behaviour $P_0(t) \sim At^{-\eta}$ with the exponent $\eta = 0.6 \pm 0.03$, to be very different to what might be expected for Lévy flights ($P_0(t) \sim t^{-2/\beta}$, that for $\beta \sim 0.6$ it implies $\eta \sim 3.33$). The slow decay in $P_0(t)$ seems to reflect the impact of an algebraic tail in the distribution of rests $\phi(t)$ between displacements.

In order to model the antagonistic interplay between scale free displacements and waiting times they use the framework of continuous time random walks (CTRW). A CTRW consists of a succession of random displacements δx_n and random waiting times δt_n each of which is drawn from a corresponding probability density function $p(\delta x_n)$ and $\phi(\delta t)$. After N iterations the position of the walker and the elapsed time are given by $x_N = \sum_n \delta x_n$ and $t_N = \sum_n \delta t_n$. The quantity of interest is the position $x(t)$ after time t and the associated probability density $W(x,t)$ which can be computed within CTRW theory. For displacements with finite variance σ^2 and waiting times with finite mean τ such a CTRW yields ordinary diffusion asymptotically, i.e. $\partial_t\, W(x,t) = D\, \partial_x^2\, W(x,t)$ with a diffusion coefficient $D = \sigma^2/\tau$.

In contrast, we assume here that both, $p(\delta x_n)$and $\phi(\delta t)$ exhibit algebraic tails, i.e. $p(\delta x_n) \sim |\delta x_n|^{-(1+\beta)}$ and $\phi(\delta t) \sim |\delta t|^{-(1+\alpha)}$, for which σ^2 and τ are infinite. In this case we can derive a bifractional diffusion equation for the dynamics of $W(x,t)$:

$$\partial_t^\alpha\, W(x,t) = D_{\alpha,\beta}\, \partial_{|x|}^\beta\, W(x,t).$$

The symbols ∂_t^α and $\partial_{|x|}^\beta$ denote fractional derivatives which are non-local and depend on the tail exponents α and β. The constant $D_{\alpha,\beta}$ is a generalised diffusion coefficient. The solution of this equation is

$$W_r(r,t) = t^{-\alpha/\beta} L_{\alpha,\beta}(r/t^{\alpha/\beta}), \tag{1.1}$$

where $L_{\alpha,\beta}$ is a universal scaling function which represents the characteristics of the process. This equation implies that the typical distance travelled scales according to $r(t) \sim t^{1/\mu}$, where $\mu = \beta/\alpha$. Thus, depending on the ratio of spatial and temporal exponents, the random walk can be effectively either superdiffusive ($\beta < 2\alpha$), subdiffusive ($\beta > 2\alpha$), or quasidiffusive ($\beta = 2\alpha$). For the exponents observed in the dispersal data ($\beta = 0.59 \pm 0.02$ and $\alpha = 0.60 \pm 0.03$) the theory predicts a temporal scaling exponent in the vicinity of unity. Therefore, dispersal remains superdiffusive despite long periods of rest.

To justify how the dispersal characteristics of bank notes carry over to the travelling behaviour of humans they observed that the power law with exponent $\beta = 0.6$ of the short time dispersal for bank notes reflects the human dispersal because the exponent remains unchanged for short time intervals of $T = 2, 4, 7$ and 14 days. Long waiting times instead may be caused by bank notes which exit the money tracking system for a long time, for instance in banks. However, the inter-report time distribution shows an exponential decay which suggests that bank notes are passed from person to person at a constant rate. Another clue comes from a comparison with two independent human travelling datasets: long distance travel on the United States aviation network and a survey on long distance travel conducted

by the United States Bureau of Transportation Statistics, which both agree well with the results of [BHG06].

In [BHG06] the authors tested the validity of the model and concluded that the dispersal of bank notes and human travelling behaviour can be described by a continuous time random walk process.

1.1.1 Lévy flights

Lévy flights (*LF*) are a particular kind of random walks. In a random walk the position is frequently defined as a sum of N independent identically distributed displacements δX_n, with *probability density function* (*pdf*) $p(\Delta x)$:

$$X_N = \sum_{n=1}^{N} \Delta X_n. \tag{1.2}$$

We start by discussing the symmetric single step pdfs in one dimension, for an ordinary random walk. According to the central limit theorem the pdf $W_Y(y, N)$ for the scaled position

$$Y_N = \frac{X_N}{\sqrt{N}} \tag{1.3}$$

is independent of N in the limit $N \to \infty$ and Gaussian, i.e.

$$\lim_{N\to\infty} W_Y(y, N) = W_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-y^2/2\sigma^2}, \tag{1.4}$$

where σ^2 is the variance of the single step δX_n. Equations (1.3) and (1.4) give the universal scaling relation

$$X_N \sim \sqrt{N}, \tag{1.5}$$

and imply that for large N the pdf $W_X(x, N)$ for the position X_N is asymptotically a spreading Gaussian:

$$W_X(x, N) \sim \frac{1}{\sqrt{N}} W_Y(x/\sqrt{N}). \tag{1.6}$$

Lévy flights belong to a class of random walks for which the central limit theorem does not apply. They can be defined by a sum of independent identically distributed random increments, but in this case the single step pdfs possess algebraic tails that make the single step second moment divergent

$$p(\Delta x) \sim \frac{1}{\Delta x^{1+\beta}} \quad 0 < \beta < 2. \tag{1.7}$$

We can then apply the Lévy Khinchin theorem, that is a generalization of the central limit theorem, and it states that, if the position of a Lévy flight is scaled according to:

$$Y_N = \frac{X_N}{N^{1/\beta}}, \tag{1.8}$$

the scaled variable possesses a pdf independent of N in the limit $N \to \infty$, i.e.

$$\lim_{N\to\infty} W_{Y,\beta}(y, N) = W_{Y,\beta}(y). \tag{1.9}$$

The limiting density $W_{Y,\beta}(y)$ is referred to as a Lévy stable law of index β and is no longer Gaussian. It can be expressed most easily in Fourier-space:

$$W_{Y,\beta}(y) = \frac{1}{2\pi}\int dk e^{-iky - D|k|^\beta}, \tag{1.10}$$

where D is some constant. Asymptotically, the limiting density has the the distribution:

$$W_{Y,\beta}(y) \sim \frac{1}{|y|^{1+\beta}}. \tag{1.11}$$

Combining Equations (1.8) and (1.10) one can obtain an explicit expression for the pdf of X_N in the limit of large step number,

$$W_{X,\beta}(x,N) \sim \frac{1}{N^{1/\beta}} W_{Y,\beta}(\frac{x}{N^{1/\beta}}). \tag{1.12}$$

This implies that the position of a Lévy flight scales superdiffusively with the step number:

$$X_N \sim N^{1/\beta} \tag{1.13}$$

1.1.2 Continuous time random walk

The continuous time random walk (CTRW), is a process defined by two pdfs: one for the spatial displacements $f(\Delta x)$ and one for random temporal increments $\phi(\Delta t)$. The CTRW then consists of pairwise random and stochastically independent events, a spatial displacement Δx and a temporal increment δt drawn from the combined pdf $p(\Delta x, \Delta t) = f(\Delta x)\phi(\Delta t)$. After N iterations the position is $X_N = \sum_{n=1}^{N} \Delta x_n$ and the time elapsed is $T_N = \sum_{n=1}^{N} \Delta t_n$. The Fourier-Laplace transforms of $W(x,t)$, pdf of the process, is given by

$$W(k,u) = \frac{1-\phi(u)}{u(1-\phi(u)f(k)))}, \tag{1.14}$$

$\phi(u)$ and $f(k)$ are the Laplace and Fourier transform of $\phi(\Delta t)$ and $f(\Delta x)$. From the inverse Laplace-Fourier transform we obtain

$$W(x,t) = \frac{1}{2\pi}\frac{1}{2\pi i}\int du \int dk e^{ut-ikx} W(k,u). \tag{1.15}$$

$W(x,t)$ may exhibit four different universal behaviours which only depend on the asymptotics of $f(\Delta x)$ and $\phi(\Delta t)$ and thus the behaviour of $f(k)$ and $\phi(u)$ for small arguments.

Ordinary Diffusion

When both, the variance of the spatial steps and the expectation value of the temporal increments exist, we have

$$f(k) = 1 - \sigma^2 k^2 + \mathcal{O}(k^4)$$
$$\phi(u) = 1 - \tau u + \mathcal{O}(u^2),$$

where σ^2 and τ are some constants. Inserting into (1.14) and using (1.15) we find that asymptotically

$$W(x,t) \sim \frac{1}{\sqrt{t}} e^{-x^2/Dt}. \tag{1.16}$$

Thus, CTRW is equivalent to Brownian motion on large spatio-temporal scales.

Lévy Flights

When the spatial displacements are drawn from a power-law pdf such as (1.7) the Fourier transform for small arguments is given by

$$f(k) = 1 - D_\beta |k|^\beta + \mathcal{O}(k^2). \tag{1.17}$$

When combined with temporal increments with finite expectation value, the same procedure as outlined above yields

$$W(x,t) \sim \frac{1}{t^{1/\beta}} L_\beta(x/t^{1/\beta}), \tag{1.18}$$

where L_β is a Lévy stable law of index β. Consequently, a CTRW with algebraically distributed spatial steps of infinite variance is equivalent to ordinary Lévy flights with a superdiffusive scaling with time $X(t) \sim t^{1/\beta}$.

Fractional Brownian motion (subdiffusion)

The complementary scenario occurs when ordinary spatial steps (finite variance and $f(k) \sim 1 - \sigma^2 k^2$) are combined with a power-law in the pdf for temporal increments,

$$\phi(\Delta t) \sim \Delta t^{-(1+\alpha)}, \ 0 < \alpha < 1. \tag{1.19}$$

In this case, the time between successive spatial increments can be very long, effectively slowing down the random walk. The Laplace transform for $\phi(\Delta t)$ is given by

$$\phi(u) = 1 - D_\alpha u^\alpha, \tag{1.20}$$

where D_α is some constant. One obtains for the position of such a random walk

$$W(x,t) = \frac{1}{2\pi} \int dk e^{-ikx} E_\alpha(-D_\alpha k^2 t^\alpha),$$

where the function E_α is the Mittag-Leffler function defined by

$$E_\alpha(z) = \sum_{n=0}^{\infty} \frac{z^n}{\Gamma(1+\alpha n)}.$$

It is easily checked that

$$W(x,t) \sim \frac{1}{t^{\alpha/2}} G_\alpha(x/t^{\alpha/2}),$$

where G_α is a non-Gaussian limiting function. From this we can obtain the scaling relation

$$X(t) \sim t^{\alpha/2}.$$

Since $\alpha < 1$ these processes are subdiffusive and sometimes referred to as fractional Brownian motions.

Ambivalent processes

The last and most interesting combination of waiting times and spatial steps is the one in which long waiting times compete and interfere with long range spatial steps, i.e. if both $\phi(\Delta t)$ and $f(\Delta x)$ decay asymptotically as a power-law, (1.7) and (1.19), so $f(k)$ and $\phi(u)$ are respectively like (1.17) and (1.20). The asymptotic pdf for the position of the ambivalent process can again be expressed in terms of a Fourier inversion and the Mittag-Leffler function according to

$$W(x,t) = \frac{1}{2\pi} \int dk e^{-ikx} E_\alpha(-D_\alpha |k|^\beta t^\alpha).$$

We can then extract the scaling relation

$$X(t) \sim t^{\alpha/\beta}.$$

The ratio of the exponents α/β resembles the interplay between subdiffusion and superdiffusion. For $\beta < 2\alpha$ the ambivalent CTRW is effectively superdiffusive, for $\beta > 2\alpha$ effectively subdiffusive. For $\beta = 2\alpha$ the process exhibits the same scaling as ordinary Brownian motion, despite the crucial difference of infinite moments and a non-Gaussian shape of the pdf $W(x,t)$.

1.2 Mobile phone

Barabási et al. [GHB08] argued that each consecutive sighting of a bank note reflects the composite motion of two or more individuals who owned the bill between two reported sightings. Thus, it is not clear whether the observed distribution reflects the motion of individual users or some previously unknown convolution between population-based heterogeneities and individual human trajectories.

To overcome this problem, they study the trajectory of 100,000 anonymized mobile phone users whose position is tracked for a six-month period. This position is known from the coordinates of the tower routing the communications, which cover an area approximately of 3 km^2. The research was performed on a random set of the total data, selected from people who made or received at least one phone call or SMS during the first and last month of the study; for this selection users who travelled outside the continental territory were excluded. As expected, data are characterized by a temporal heterogeneity, but the authors checked that it does not affect the results on the observed travel patterns.

To explore the statistical properties of the population's mobility patterns, they measured the distance between user's positions at consecutive calls, and found that the distribution of displacements over all users is well approximated by a truncated power-law:

$$P(\Delta r) = (\Delta r - \Delta r_0)^{-\beta} \exp(-\Delta r/\kappa) \tag{1.21}$$

with exponent $\beta = 1.75 \pm 0.15$, $\Delta r_0 = 0.5km$ and cutoff values $\kappa|_{D_1} \equiv 400km$ and $\kappa|_{D_2} \equiv 80km$. The observed scaling exponent is not far from $\beta = 1.59$ ([GHB08, p. 799]), observed for bank note dispersal, suggesting that the two distributions may capture the same fundamental mechanism driving human mobility patterns.

Equation (1.21) is compatible with the hypothesis that human motion follows a *truncated* Lévy flight (*TLF*), but the observed shape of $P(\Delta r)$ could also be explained by a population-based heterogeneity, corresponding to the inherent differences between individuals, coexisting with individual Lévy trajectories.

In the description of individual trajectories, Barabási et al. characterized the linear size occupied by each user's trajectory up to time t by its *radius of gyration* defined as:

$$r_a^g(t) = \sqrt{\frac{1}{n_c^a(t)} \sum_{i=1}^{n_c^a} (\vec{r_i^a} - \vec{r_{cm}^a})^2},$$

where $\vec{r_i^a}$ represents the $i = 1, ..., n_a^c(t)$ positions recorded for user a and $\vec{r_{cm}^a} = 1/n_c^a(t) \sum_{i=1}^{n_c^a} \vec{r_i^a}$ is the center of mass of the trajectory. The radius of gyration distribution $P(r_g)$ can be approximated with a truncated power-law:

$$P(r_g) = (r_g + r_g^0)^{-\beta_r} \exp(-r_g/\kappa) \tag{1.22}$$

with $r_g^0 \equiv 5.8km$, $\beta_r = 1.65 \pm 0.15$ and $\kappa = 350km$. An ensemble of Lévy agents displays a significant degree of heterogeneity in r_g; however, this is not sufficient to explain the truncated power-law distribution $P(r_g)$ exhibited by the mobile phone users: the difference in the range of typical mobility patterns of individuals (r_g) has a strong impact on the truncated Lévy behaviour.

If individual trajectories are described by an LF or TLF, then the radius of gyration should increase with time as $r_g(t) \sim t^{3/(2+\beta)}$, whereas, for an RW, $r_g(t) \sim t^{1/2}$; that is, the longer we observe a user, the higher the chance that she/he will travel to areas not visited before. Data results indicate that the time dependence of the average radius of gyration of mobile phone users is better approximated by a logarithmic increase, not only a manifestly slower dependence than the one predicted by a power law but also one that may appear similar to a saturation process.

Users with small r_g travel mostly over small distances, whereas those with large r_g tend to display a combination of many small and a few larger jump sizes; so rescaling the distributions with r_g, the data collapse into a single curve. This suggests that a single jump size distribution characterises all users, regardless of their r_g. $P(\Delta r|r_g) \sim r_g^{-\alpha} F(\Delta r/r_g)$ is then an r_g-independent function with asymptotic behaviour.

The authors measured the probability that a user returns to the position where he/she was first observed after t hours $F_{pt}(t)$: for a two-dimensional random walk, $F_{pt}(t) \sim 1/(t \ln^2(t))$. In contrast, in this case the return probability is characterized by several peaks at 24 h, 48 h and 72 h, capturing a strong tendency of humans to return to locations they visited before, describing the recurrence and temporal periodicity inherent to human mobility.

An important quantity for modelling human mobility patterns is the probability density function $\tilde{\Phi}_a(x, y)$ to find an individual a in a given position (x, y). In order to compare different users' trajectories, in [GHB08] the authors chose for them a common reference frame calculated a posteriori. As in the mechanics of rigid bodies, every (two dimensional) trajectory is characterized by a 2×2 matrix known as the *tensor of inertia* I, where the number of times a user visited a given location is the mass associated with that particular

position.

$$I = \begin{pmatrix} I_{xx} & I_{xy} \\ I_{yx} & I_{yy} \end{pmatrix},$$

if we denote a user's trajectory with a set of locations $(x_1, y_1), (x_2, y_2), \ldots (x_{n_c}, y_{n_c})$, where n_c is the number of positions available for the user, we have

$$I_{xx} \equiv \sum_{i=1}^{n_c} y_i^2,\ I_{yy} \equiv \sum_{i=1}^{n_c} x_i^2,\ I_{xy} = I_{yx} \equiv -\sum_{i=1}^{n_c} x_i y_1.$$

Since the tensor I is symmetric, it is possible to find a set of coordinates in which it will be diagonal. The corresponding eigenvectors determine the principal axes ($\hat{e}_1$ and $\hat{e}_2$), representing the symmetry axes of a given trajectory. They transformed each user's principal axes ($\hat{e}_1$, $\hat{e}_2$) to a common intrinsic reference frame ($\hat{e}_x$, $\hat{e}_y$). They discovered that the larger an individual's r_g, the more pronounced is his anisotropy. So they scaled the trajectories on the intrinsic axes with the standard deviation of the locations for each user a:

$$\sigma_x^a = \sqrt{\frac{1}{n_c^a} \sum_{i=1}^{n_c^a} (x_i^a - x_{cm}^a)^2}.$$

After scaling, the shapes of the trajectories look similar, despite the fact that we are showing users with significantly different mobility patterns and ranges. This is the underlying procedure that allows them to obtain an universal density function $\tilde{\Phi}(x/\sigma_x, y/\sigma_y)$.

They find that, in contrast with the random trajectories predicted by the prevailing Lévy flight and random walk models, human trajectories show a high degree of temporal and spatial regularity, each individual being characterized by a time-independent characteristic travel distance and a significant probability to return to a few highly frequented locations. After correcting for differences in travel distances and the inherent anisotropy of each trajectory, the individual travel patterns collapse into a single spatial probability distribution, indicating that, despite the diversity of their travel history, humans follow simple reproducible patterns. This inherent similarity in travel patterns could impact all phenomena driven by human mobility, from epidemic prevention to emergency response, urban planning and agent-based modelling.

Taken together, their results suggest that the Lévy statistics observed in bank note measurements capture a convolution of the population heterogeneity shown in equation (1.22) and the motion of individual users. Individuals display significant regularity, because they return to a few highly frequented locations, such as home or work. This regularity does not apply to the bank notes: a bill always follows the trajectory of its current owner; that is, dollar bills diffuse, but humans do not. Contrary to bank notes, mobile phones are carried by the same individual during his/her daily routine, offering the best proxy to capture individual human trajectories.

1.3 Exploration and preferential return

Both dollar-bill tracking and mobile-phone data indicate that the aggregated jump-size (Δr) and waiting-time (Δt) distributions characterizing human trajectories are fat-tailed,

that is,

$$P(\Delta r) \sim |\Delta r|^{-1-\alpha}, \text{ with} 0 < \alpha \leq 2 \tag{1.23}$$

and

$$P(\Delta t) \sim |\Delta t|^{-1-\beta}, \text{ with} 0 < \beta \leq 1 \tag{1.24}$$

where r denotes the distances covered by an individual between consecutive sightings and t is the time spent by an individual at the same location. These findings suggest that if we look at Δr, human trajectories are best described as Lévy flights or CTRWs. In the last section we saw that, in contrast with the random trajectories predicted by the prevailing Lévy flights and random walk models, human trajectories show a high degree of temporal and spatial regularity, each individual being characterized by a time-independent characteristic travel distance and a significant probability to return to a few highly frequented locations.

Barabási et al. [SKWB10] studied trajectories of three million anonymized mobile-phone users and found the same behaviour of (1.23) and (1.24) with exponent $\alpha = 0.55 \pm 0.05$ and cutoff $\Delta r \sim 100$ km, corresponding to the distance people could reasonably cover in an hour; and $\beta = 0.8 \pm 0.1$ and cutoff $\Delta t \sim 17$h , probably capturing the typical awake period of an individual. Although this seems to be in agreement with CTRWs, they found some contradictions:

- the number of distinct locations $S(t)$ visited by a randomly moving object is expected to follow

$$S(t) \sim t^{\mu}, \tag{1.25}$$

  with $\mu = \beta$. They find $\mu = 0.6 \pm 0.02$, smaller than expected.
- Visitation frequency: the probability f of a user to visit a given location is expected to be asymptotically (t approach infinity) uniform everywhere for both Lévy flights and CTRWs. In contrast, the visitation patterns of humans is rather uneven, so that the frequency f of the k^{th} most visited location follows Zipf's law

$$f_k \sim k^{-\zeta}$$

  where $\zeta \sim 1.2 \pm 0.1$. This suggests that the visitation frequency distribution follows $P(f) = f^{-(1+1/\zeta)}$.
- Ultraslow diffusion: the CTRW model predicts that the mean square displacement (MSD) asymptotically follows $\langle x^2(t) \rangle \sim t^{\nu}$ with $\nu = 2\beta/\alpha \sim 3.1$. As both $P(\Delta r)$ and $P(\Delta t)$ have cutoffs, asymptotically the MSD should converge to a Brownian behaviour with $\nu = 1$. In other words, CTRW predicts that the longer we follow a human trajectory, the further it will drift from its initial position. Yet, humans have a tendency to return home on a daily basis, suggesting that simple diffusive processes, which are not recurrent in two dimensions, do not offer a suitable description of human mobility.

Barabási et al. suggest that two generic mechanisms, exploration and preferential return, both unique to human mobility, are missing from the traditional random-walk (Lévy flight or CTRW) models:

- Exploration: with probability

$$P_{new} = \rho S^{-\gamma}$$

the individual moves to a new location (different from the S locations he/she visited before). The distance Δr that he/she covers during this exploratory jump is chosen from the $P(\Delta r)$ distribution and his/her direction is selected to be random. As the individual moves to this new position, the number of previously visited locations increases from S to $S+1$.

- Preferential return: with the complementary probability

$$P_{ret} = 1 - \rho S^{-\gamma}$$

the individual returns to one of the S previously visited locations. In this case, the probability π_i of visiting location i is chosen to be proportional to the number of visits the user previously had to that location. That is, we assume that

$$\pi_i = f_i. \tag{1.26}$$

With this assumption it is possible to solve the problem of the previous models.

- The probability that an individual moves to a new location is proportional to $S^{-\gamma}$, that is, $dS/dn \propto S^{-\gamma}$, predicting $S \sim n^{1/(1+\gamma)}$, where n is the total number of discrete moves the individual had up to time t . For a fat-tailed waiting-time distribution $P(\Delta t) \sim |\Delta t|^{-1-\beta}$ the time t scales with the number of jumps n as $t \sim n^{1/\beta}$, showing that $S(t)$ follows (1.25) with the exponent $\mu = \beta/(1+\gamma)$. Data are in agreement with this prediction.
- We notice that m_i , the number of visits to location i, increases like $dm_i/dn = \pi_i(1 - P_{new})$, where $\pi_i = f_i = m_i/\sum_i m_i(n)$ is the probability of returning to the location i during step n. When $\gamma > 0$, in the limit of $S(t) \to \infty$ the probability of exploring a new location is negligible compared with the return visits; thus, asymptotically we have $dm_i/dn = m_i/\sum_i m_i(n)$. As $\sum_i m_i(n) = n$, we obtain $m_i(n) = n/n_i$, where n_i denotes the jump during which location i was first visited, at which moment $m_i(n_i) = 1$. Owing to preferential return (1.26), the earlier a location is visited, the more it is visited later. Thus, the ranking k_i for location i coincides with the order in which it was first visited, that is, $k_i = S(n_i) \sim n_i^{(1+\gamma)}$. As the visitation frequency f_i is proportional to $m_i(n) = n/n_i$, we have $f_k \sim k^{-\zeta}$ with the exponent $\zeta = 1 + \gamma$. In general, they find $\zeta = 1 + \gamma$, if $\gamma > 0$, or $\zeta = 1 - \rho$, if $\gamma = 0$.
- This model predicts

$$\langle \Delta x^2 \rangle^{\alpha/2} \sim \log\left(\frac{1 - S^{1-\zeta}}{\zeta - 1}\right) + \text{const.} \tag{1.27}$$

Another interesting aspect that they found is population heterogeneity: the radius of gyration r_g of the trajectory of different individuals is found to follow a fat-tailed distribution. Their model can reproduce this feature as well, indicating that the fat-tailed $P(r_g)$ is a consequence of the inherent fluctuations present within the model and it is rooted in the $P(\Delta r)$ distribution.

It is important to note that in contrast with the traditional random-walk, Lévy flight or CTRW models, this model is dynamically quenched. That is, after an individual explores a new location, he/she will have an increasing tendency to return to it in the future, generating a recurrent and relatively stable mobility pattern for each individual.

This model is designed to capture the long-term spatial and temporal scaling patterns; thus, in its present from, it does not reproduce the short-term temporal order and correlations potentially present in individual mobility. This is important also for practical considerations: many human-mobility-driven processes, from epidemic spreading to city planning, are driven by the asymptotic characteristics of human mobility.

Chapter 2

Models of spatial flows

A spatial interaction is a realised movement of people, freight or information between an origin and a destination. It is a transport demand/supply relationship expressed over a geographical space. Spatial interactions cover a wide variety of movements such as journeys to work, migrations, tourism, the usage of public facilities, the transmission of information or capital, the market areas of retailing activities, international trade and freight distribution.

Each spatial interaction, as an analogy for a set of movements, is composed of an origin/destination pair. Each pair can itself be represented as a cell in a matrix where rows are related to the locations of origin, while columns are related to locations of destination. Such a matrix is commonly known as an *Origin/Destination matrix* (*OD matrix*), or a *spatial interaction matrix*.

The basic assumption concerning many spatial interaction models is that flows are a function of the attributes of the locations of origin, the attributes of the locations of destination and the friction of distance between the concerned origins and destinations.

The *gravity model* is one of basic models of spatial analysis in geography and social physics. It provides an empirically effective approach to modeling spatial interaction. The model is originally proposed to describe population migration between two regions (Carey, 1858, [Car65]; Grigg, 1977, [Gri]; Ravenstein, 1885, [Rav85]). Afterward, it is employed to measure the force of attraction between any two geographical objects such as cities, firms, and retail stores. Today, the model can be found in many subjects like economics and sociology. A lot of variants of the model came out, and different forms of gravity models have different spheres of application (Erlander, 1980, [Erl80]; Haynes and Fotheringham, 1984, [HF84]; Sen and Smith, 1995, [SS95]).

2.1 Gravity models

The gravity model is one of the most important spatial interaction methods. It is named as such because it uses a similar formulation to Newton's gravitation model. Accordingly, the attraction between two objects is proportional to their mass and inversely proportional to the square of their respective distance. Consequently, the general formulation of spatial interactions can be adapted to reflect this basic assumption to form the elementary formulation

of the gravity model:

$$T_{ij} = k\frac{P_i P_j}{d_{ij}^2} \tag{2.1}$$

- P_i and P_j are the weights (e.g. population) of the location of origin i and the location of destination j.
- d_{ij} is the distance, or any measure related to the friction of space, between the locationof origin and the location of destination.
- k is a proportionality constant related to the temporal rate of the event being measured. For instance, if the same system of spatial interactions is considered, the value of k will be higher if interactions are considered for a year instead of a week.

Thus, spatial interactions between locations i and j are proportional to their respective weights divided by their distance.

The gravity model can be extended to include several parameters:

$$T_{ij} = k\frac{P_i^\alpha P_j^\beta}{d_{ij}^\gamma} \tag{2.2}$$

- α is the potential to generate movements (emissiveness). For movements of people, α is often related to an overall level of welfare. For instance, it is logical to infer that for retailing flows, a location having higher income levels will generate more movements.
- β is the potential to attract movements (attractiveness). Related to the nature of economic activities at the destination. For instance, a center having important commercial activities will attract more movements.
- γ is a parameter of transport friction related to the efficiency of the transport system between two locations. This friction is rarely linear as the further the movement the greater the friction of distance. For instance, a highway between two locations will have a weaker beta index than a road.

Different kinds of the Gravity Model may consider different kind of weights. Another variation is to consider different kind of functions to descibed the dependence on the distance; for example, a negative exponential instead of a power-law. The function of the distance is usually called the *deterrence function*. In the next chapter we will use different formulations of the model to fit data, and then check which of these is the most appropriate model.

In the formulation of the Gravity Model we just described, an overall increase in the "attractiveness" of the various locations will have the effect of increasing the total number of trips made by people. Another formulation of the gravity model can be to to take the total number of trips as already given. This is equivalent to adding a constraint equation

$$T_i = \sum_j T_{ij}$$

that leads to

$$T_{ij} = T_i \frac{P_j^\beta}{d_{ij}^\gamma} \frac{1}{\sum_j \frac{P_j^\beta}{d_{ij}^\gamma}}. \tag{2.3}$$

This is called *Total Interaction Constrained Gravity Model*. If we fixed the number of trips from each location, we call that model *Singly-costrained gravity model*. If both total trips

from and to every location are fixed, we have the *Doubly-costrained gravity model*. In the next section we show a derivation of this model with argument of maximum entropy.

2.2 Maximum entropy argument

Here we present the so-called *Maximum entropy argument* f[Wil69], that is a statistical derivation which constitutes a theoretical base for (double-constraint) gravity models. The basic assumption of Maximum entropy argument is that the probability of the distribution of flows is proportional to the number of states of the system which give rise to this particular distribution, and which satisfied some costraints.

Let us examine a single trip purpose, like the journey to work. We consider for simplicity only one mode of transport and one type of traveller. Suppose the region is divided into zones and that T_{ij} is the number of trips between zones i and j, c_{ij} is the cost (which takes into account for money, travel time, etc.) of travelling between i and j, O_i is the total number of trip origins at i and D_j is the total number of trip destination at j. A general formulation of gravity model, used to estimate the number of trips between i and j, T_{ij}, in terms of the other variables, can be written as:

$$T_{ij} = A_i B_j O_i D_j f(c_{ij}) \tag{2.4}$$

where f is some decreasing function cost. A_i and B_i are defined as:

$$A_i = \big[\sum_j B_j D_j f(c_{ij})\big]^{-1} \tag{2.5}$$

$$B_j = \big[\sum_i A_i O_i f(c_{ij})\big]^{-1} \tag{2.6}$$

in order to satisfy the constraint equations

$$\sum_j T_{ij} = O_i \tag{2.7}$$

and

$$\sum_i T_{ij} = D_j. \tag{2.8}$$

(2.5) and (2.6) are usually solved by some iterative procedure.

We start from this formulation and assume another constraint equation to fix the total amount C spent on these trips in the region at the given point in time:

$$\sum_i \sum_j T_{ij} c_{ij} = C. \tag{2.9}$$

We write $T = \sum_i O_i = \sum_j D_j$ for the total number of trips.

If we call $w(T_{ij})$ the number of distinct arrangements of individuals which give rise to the distribution T_{ij}, corresponding to the number of ways in which T_{11} can be selected from T, T_{12} from $T - T_{11}$ and so on; thus we have

$$w(T_{ij}) = \frac{T!}{T_{11}!(T-T_{11})!}\frac{(T-T_{11})!}{T_{12}!(T-T_{11}-T_{12})!}\cdots = \frac{T!}{\prod_{ij} T_{ij}!}. \tag{2.10}$$

The total number of possible states is then

$$W = \sum w(T_{ij}) \tag{2.11}$$

where the summation is over all the distributions T_{ij} which satisfy the constraints (2.7)-(2.9). However, the maximum value of $w(T_{ij})$ turns out to dominate the other terms of the sum to such an extent that the distribution T_{ij} which give rise to this maximum is far the most probable distribution. Now we look for this maximum. In order to obtain $\{\tilde{T}_{ij}\}$ which maximises $w(T_{ij})$ with the constraints (2.7)-(2.9) we the Lagrangian multipliers method for function M:

$$M = \log w + \sum_i \lambda_i^{(1)}(O_i - \sum_j T_{ij}) + \sum_j \lambda_j^{(2)}(D_j - \sum_i T_{ij}) + \beta(C - \sum_{ij} T_{ij})$$

$\lambda_i^{(1)}$, $\lambda_j^{(2)}$, and β are Lagrangian multipliers, and the equation to be solved is

$$\frac{\partial M}{\partial T_{ij}}(\tilde{T}_{ij}) = 0.$$

We maximise $\log w$ rather than w, so we can use Stirling's approximation $\log N! = N \log N - N$ to estimate the factorial terms; this give us:

$$\frac{\partial \log N!}{\partial N} = \log N$$

and so

$$\frac{\partial M}{\partial T_{ij}} = -\log T_{ij} - \lambda_i^{(1)} - \lambda_j^{(2)} - \beta c_{ij}.$$

The maximum is therefore

$$\tilde{T}_{ij} = \exp\left[-\lambda_i^{(1)} - \lambda_j^{(2)} - \beta c_{ij}\right].$$

Substituting in (2.7) and (2.8) we obtain $\lambda_i^{(1)}$ and $\lambda_j^{(2)}$:

$$e^{-\lambda_i^{(1)}} = \frac{O_i}{\sum_j e^{-\lambda_j^{(1)} - \beta c_{ij}}} \tag{2.12}$$

$$e^{-\lambda_j^{(2)}} = \frac{D_j}{\sum_i e^{-\lambda_i^{(1)} - \beta c_{ij}}}. \tag{2.13}$$

To obtain the result in more familiar form, write

$$A_i = e^{-\lambda_i^{(1)}}/O_i \tag{2.14}$$

$$B_j = e^{-\lambda_j^{(2)}}/D_j \tag{2.15}$$

and then

$$T_{ij} = A_i B_j O_i D_j e^{-\beta c_{ij}} \tag{2.16}$$

where, using equations (2.12)-(2.15),

$$A_i = \left[\sum_j B_j D_j e^{-\beta c_{ij}}\right]^{-1} \tag{2.17}$$

$$B_j = \left[\sum_i A_i O_i e^{-\beta c_{ij}}\right]^{-1}. \tag{2.18}$$

Hence the most probable distribution of trips is the gravity model (in its doubly-costrained form) discussed earlier. The meaning of what we have shown is that, given the total number of trip origins and destinationas for each location, given the costs of travelling between each pair of zones, and fixed the total cost spent in the region at the given point in time, then there is a most probable distribution of trips, which is gravity model.

The most important problem at this point is how to express the cost. In order to recover a power law distribution, one needs a logarithmic dependence on distance: $c_{ij} = a \log(r_{ij})$ which leads to $T_{ij} \propto r_{ij}^{-\beta a}$. If the cost is proportional to distance, the number of trips decays exponentially with distance. We thus recover two of the most important forms used in empirical studies and in model, but the exact form of the cost dependence with distance remains unsolved.

There is a long discussion about the validity of this approach in [ES90] but we note that it assumes in particular that all individuals act independently from each other. This is obviously not correct when we introduce congestion which induces correlations between individuals. In such conditions, it is clear that individual choices are correlated and that this entropy maximization can give reasonable results in the limit of small traffic only.

2.3 Intervening opportunities model

The basic idea behind the *intervening-opportunities model* is that trip making is not explicitly related to distance but to the relative accessibility of opportunities for satisfying the objective of the trip. The original proponent of this approach was *Stouffer*, 1940, ([Sto40]) who applied this approach to migration patterns between services and residences. The theory was further developed by Schneider, 1959, ([Sch59]) to the general framework that is used today.

The law of intervening opportunities as proposed by Stouffer states "The number of persons going a given distance is directly proportional to the number of opportunities at that distance and inversely proportional to the number of intervening opportunities.". An *opportunity* is a destination that a trip-maker considers as a possible termination point for their journey and an *intervening opportunity* is an opportunity that is closer to the trip maker than the final destination but is rejected by the trip-maker.

This hypothesis may be expressed as:

$$T_{ij} = k \frac{A_j}{V_j} \tag{2.19}$$

where A_j is the total number of destination opportunities in zone j and V_j is the number of intervening destination opportunities between zones i and j, k is a proportionality constant.

Schneider proposed a modified Stouffer hypothesis: "The probability that a trip will terminate in some volume of destination points is equal to the probability that this volume contains an acceptable destination times the probability that an acceptable destination closer to the origin of the trips has not been found.". This was represented mathematically by Ruiter

[Rui67] as

$$dP = L[1 - P(V)]dV \tag{2.20}$$

where dP is the probability that a trip will terminate when considering dV possible destinations; the 'subtended volume' V is the cumulative total number of destination opportunities considered up to the destination being considered; dV is an element of the subtended volume at the surface of the volume; $P(V)$ represents the opportunity that a trip terminates when V destinations are considered; L is a constant probability of a possible destination being accepted if it is considered. The solution of equation (2.20) is

$$P(V) = 1 - \exp(-LV). \tag{2.21}$$

With the expected trip-interchange, T_{ij}, we get:

$$T_{ij} = O_i[P(V_{j+1}) - P(V_j)], \tag{2.22}$$

where O_i represents the total number of opportunities at location i. In [Eas84] Eash showed that the gravity and IO models are "fundamentally the same" and are both derivable from entropy maximization theory. Eash also noted that the difference is how the "cost" of travel is considered. Although the gravity model considers this "cost" as a function of the distance, the opportunity model considers the "cost" as the difficulty to satisfy a trip's purpose. The gravity model then treats the distance variable as a continuous cardinal variable, whilst the opportunity model treats the distance as an ordinal variable.

2.4 Radiation Model

Despite its widespread use, the gravity law has notable limitations: lack of a rigorous derivation (entropy maximization fail to offer the functional form of $f(r)$), inability to predict mobility in region where we lack systematic traffic data, systematic predictive discrepancies, analytical inconsistency (it predicts that the number of commuters increases without limit as we increase the destination population, but obviously commuters cannot exceed the origin population), being deterministic it cannot account for fluctuations.

The *Radiation Model* is a proposal to solve these problems. We present it in the context of commuting flows. Whereas commuting is a daily process, its source and destination is determined by job selection. The *Radiation model* assume that the selection consist on two step. First, an individual look for a job from all counties: the number of employment opportunities is proportional to resident population; the benefits of a potential job is described by a number z, randomly chosen from a distribution $p(z)$, where z represents a combination of salary, working hours, conditions, and others. The second step for the individual is to choose the closest job to his/her home, whose benefits z are higher than the best offer available in his/her home county.

This process applied in proportion to the resident population in each couty, determines the daily commuting flow across the country. The model has three parameters, the benefit distribution $p(z)$, the job density and the total number of commuters, but it turn out that none of them affect the flux distribution, making the model parameter-free. It is called

Radiation model because it can be formulated in terms of radiation and absorption processes, in analogy with Physics. Imagine the location of origin, i, as a source emitting an outgoing flux of identical and independent units (particles). We define the emission/absorption process through the following two steps:

- We associate to every particle, X, emitted from location i a number, $z_X^{(i)}$, that represents the absorption threshold for that particle. A particle with large threshold is less likely to be absorbed. We define $z_X^{(i)}$ as the maximum number obtained after m_i random extractions from a preselected distribution, $p(z)$ (m_i is the population in location i). Thus, on average, particles emitted from a highly populated location have a higher absorption threshold than those emitted from a scarcely populated location. We will show below that the particular choice of $p(z)$ do not affect the final results.
- The surrounding locations have a certain probability to absorb particle X: $z_X^{(j)}$ represents the absorbance of location j for particle X, and it is defined as the maximum of n_j extractions from $p(z)$ (n_j is the population in j). The particle is absorbed by the closest location whose absorbance is greater than its absorption threshold.

By repeating this process for all emitted particles we obtain the fluxes across the entire country. We can calculate the probability of one emission/absorption event between any two locations, and thus obtain an analytical prediction for the flux between them. Let $P(1\,|\,m_i, n_j, s_{ij})$ be the probability that a particle emitted from location i with population m_i is absorbed in location j with population n_j, given that s_{ij} is the total population in all locations (except i and j) within a circle of radius r_{ij} centered at i (r_{ij} is the distance between i and j). According to the radiation model, we have

$$P(1\,|\,m_i, n_j, s_{ij}) = \int_0^\infty dz P_{m_i}(z) P_{s_{ij}}(< z) P_{n_j}(> z) \tag{2.23}$$

where $P_{m_i}(z)$ is the probability that the maximum value extracted from $p(z)$ after m_i trials is equal to z:

$$P_{m_i}(z) = \frac{dP_{m_i}(< z)}{dz} = m_i p(< z)^{m_i - 1} \frac{dp(< z)}{dz}. \tag{2.24}$$

Similarly $P_{s_{ij}}(< z) = p(< z)^{s_{ij}}$ is the probability that s_{ij} numbers extracted from the $p(z)$ distribution are all less than z; and $P_{n_j}(> z) = 1 - p(< z)^{n_j}$ is the probability that among n_j numbers extracted from $p(z)$ at least one is greater than z. Thus (2.23) represents the probability that one particle emitted from a location with population m_i is not absorbed by the closest locations with total population s_{ij}, and is absorbed in the next location with population n_j. After evaluating the above integral, we obtain

$$\begin{aligned} P(1\,|\,m_i, n_j, s_{ij}) &= m_i \int_0^\infty dz \frac{dP(< z)}{dz} \left[p(< z)^{m_i + s_{ij} - 1} - p(< z)^{m_i + n_j + s_{ij} - 1} \right] \\ &= m_i \left[\frac{1}{m_i + s_{ij}} - \frac{1}{m_i + n_j + s_{ij}} \right] \\ &= \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \end{aligned} \tag{2.25}$$

which is indipendent of the distribution $p(z)$ and is invariant under rescaling of the population by the same multiplicative factor (n_{jobs}). The probability $P(T_{i1}, T_{i2}, \ldots, T_{iL})$ for a particular

sequence of absorptions, $(T_{i1}, T_{i2}, \ldots, T_{iL})$, of the particles emitted at location i is given by the multinomial distribution:

$$P(T_{i1}, T_{i2}, \ldots, T_{iL}) = \prod_{j \neq i} \frac{T_i!}{T_{ij}!} p_{ij}^{T_{ij}} \quad with \quad \sum_{j \neq i} T_{ij} = T_i \tag{2.26}$$

where T_i is the total number of particles emitted by location i, and $p_{ij} \equiv P(1|m_i, n_j, s_{ij})$. The distribution (2.26) is normalized. The probability that exactly T_{ij} particles emitted from location i are absorbed in location j is obtained by marginalizing probability (2.26) :

$$\begin{aligned} P(T_{ij} \,|\, m_i, n_j, s_{ij}) &= \sum_{\{T_{ik} | k \neq i,j; \sum_{k \neq i} T_{ik} = T_i\}} P(T_{i1}, T_{i2}, \ldots, T_{ij}, \ldots, T_{iL}) \\ &= \frac{T_i!}{T_{ij}!(T_i - T_{ij})!} p_{ij}^{T_{ij}} (1 - p_{ij})^{T_i - T_{ij}} \end{aligned} \tag{2.27}$$

that is a binomial distribution with average

$$\langle T_{ij} \rangle \equiv T_i p_{ij} = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})} \tag{2.28}$$

and variance $T_i p_{ij}(1 - p_{ij})$.

(2.28) represent the fundamental equation of the radiation model. It solves some limitations of the gravity law: it has a rigorous derivation and has no free parameters. It has been shown [SGMB12] that it has better performance with migration flows in the USA.

The radiation model might provide further insights on the problem of defining human agglomerations, that is an important issue in the study of cities because different definitions affect conclusions regarding the statistical distribution of urban activity.

Chapter 3

Fitting Gravity Model: Generalized Linear Models

In this chapter we present a useful way to fit the Gravity Model, based on *Generalized Linear Models* (GLM), which is a generalization of linear regression and is better suited in this context, because it takes into account the fluctuations in a more proper way.

In the end we apply this technique to relocation flows in USA, comparing the goodness of fit with different formulation of Gravity Models (different deterrence functions, different constraints ...).

With this technique it is necessary to decide from the beginning which function to use as the deterrence function. In the next chapter we will show a method of fitting the Gravity Model without the need to decide a priori which function to use.

3.1 Fitting Gravity Model

Flowerdew et al. [FA82] [1] suggest a method for fitting the gravity model, in the form

$$T_{ij} = k\frac{P_i^\alpha P_j^\beta}{d_{ij}^\gamma}. \tag{3.1}$$

If we replace T_{ij} with his mean μ_{ij} and we take the logarithm, we obtain

$$\log \mu_{ij} = \log k + \alpha \log P_i + \beta \log P_j - \gamma \log d_{ij}. \tag{3.2}$$

It was common to use a *log-normal model* to estimate the values of the parameters $\log k$, α, β, γ, with an ordinary least-squares (OLS) multiple regression analysis, which finds the estimates by minimize the sum of squared residuals. It uses $\log n_{ij}$, the logarithm of the number of migrants moving from i to j, and assumes that:

$$\log \mu_{ij} = \log k + \alpha \log P_i + \beta \log P_j - \gamma \log d_{ij} + u_{ij}, \tag{3.3}$$

with u_{ij} indipendent random variables normally distribuited (with zero mean and identical variance).

Log-normal models presents several weakness:

[1] They used a data set consist of observations on one year migration flow between the 126 SMLA's (Standard Metropolitam Labor Areas) for Great Britain.

- the regression estimates the logarithms of μ_{ij}, so the antilogarithm of these quantities are biased estimate of μ_{ij}. As a consequence, large flow are understimated, and the sum of the estimated flows is considerably less than the sum of the observed flows.
- F_{ij} are supposed to be log-normally distribuited around the estimated, leading to negative values of F_ij.
- The assumption of identical variances for the u_{ij} implies that the expected difference between the estimate of $\log \mu_{ij}$ and $\log n_{ij}$ is the same for every pairs of locations (that is, there is the same probability for an observed flow of two and an estimated of one, as for a flow of 200 in relation of an estimate of 100!)
- When flow are zero, a small positive number should be added to all observation, and this can affect the predicted parameters.

To overcome these problems, we should notice that movements can be considered indipendent, characterized by a small constant probability for a person in i to move in j, and the population of i is large. This implies that the probability that k people move will follow a Poisson distribution:

$$P(n_{ij} = k) = \frac{e^{-\mu_{ij}} \mu_{ij}^k}{k!}, \tag{3.4}$$

and the mean μ_{ij} will follow (3.2), so:

$$\mu_{ij} = exp(\log k + \alpha \log P_i + \beta \log P_j - \gamma \log d_{ij}). \tag{3.5}$$

In this case, difference between μ_{ij} and the observed flows are the result of the particolar realization of the Poisson process, and this difference is measured on the scale of n_{ij} and not of his logarithm. Another difference with log-normal model is that now the variance is equal to the mean, and so it is not constant.

Differently from log-normal model, estimates derived from Poisson model are usually of the same order of magnitude as observed flows and the sum of observed and estimate flows are approximately equal.

Poisson model appears [FA82] to give a better description to data than the log-normal model.

3.2 Generalized Linear Model Theory

Let $y_1, \ldots, y_n$ denote n independent realization of a random variable Y_i . In the general linear model the assumption are that Y_i is normally distributed with mean μ_i and variance σ^2:

$$Y_i \sim N(\mu_i, \sigma^2),$$

where the expected value μ_i is a linear function of p predictors that take values $\boldsymbol{x_i} = (x_{i1}, \ldots, x_{ip})$, and $\boldsymbol{\beta}$ is a vector of unknown parameters:

$$\mu_i = \boldsymbol{x_i}\boldsymbol{\beta}. \tag{3.6}$$

The Generalized Linear Model, formulated by Nelder and Wedderburn (1972) [NW72] is a generalization of general linear model for the case of not normally distribuited random variable.

The first element is the assumption that the observations come from a distribution in the exponential family, i.e. $Y_i \sim f(\theta, \phi)$, where

$$f(y_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right\}, \tag{3.7}$$

θ_i and ϕ are parameters (θ_i is the parameter of interest, as ϕ is a nuisance parameter (as σ in regression); $b(\theta_i)$, $a_i(\phi)$ and $c(y_i, \phi)$ are known functions. The function $a_i(\phi)$ has the form $a_i(\phi) = \phi/p_i$, where ϕ, called the dispersion parameter, is constant over observations, and p_i is a known prior weight, usually 1, that varies from observation to observation. It can be shown that if Y_i has a distribution in the exponential family then [2]

$$\begin{aligned} 0 &= E\left\{\frac{Y - b'(\theta)}{a(\psi)}\right\} \\ E\left\{\frac{-b''(\theta)}{a(\psi)}\right\} &= -E\left\{\frac{Y - b'(\theta)}{a(\psi)}\right\}^2 \end{aligned} \tag{3.8}$$

so it has the following mean and variance:

$$\begin{aligned} E(Y_i) &= \mu_i = b'(\theta_i) \\ var(Y_i) &= \sigma_i^2 = b''(\theta_i)a_i(\phi), \end{aligned} \tag{3.9}$$

where $b'(\theta_i)$ and $b''(\theta_i)$ are the first and second derivatives of $b(\theta_i)$. The expectation of Y only depends on θ whereas the variance of Y depends on θ and ψ. All the commonest distributions, like the normal, binomial, Poisson, exponential, gamma and inverse Gaussian, are special cases of exponential family.

The second important generalization is that, instead of the mean, a transformed mean

$$\eta_i = g(\mu_i)$$

follows a linear model

$$\eta_i = \boldsymbol{x_i\beta}. \tag{3.10}$$

The function $g(\mu_i)$ is called the *link function* is a one-to-one continuous differentiable transformation; η_i is called the *linear predictor*. In order to obtain μ_i we simply need to invert the link function

$$\mu_i = g^{-1}(\boldsymbol{x_i\beta});$$

the reason of this transformation is that usually the model for μ_i is more complicated than the model for η_i. It is important to notice that the transformation is not of the response y_i, but it is of his expected value μ_i.

Examples of commonly used link functions are the identity, log, reciprocal, logit and probit. When the link function makes the linear predictor η_i the same as the canonical parameter θ_i, we call it *canonical link*. For example, the identity is the canonical link for the normal distribution. The advantage of the canonical link is that all the information about $\boldsymbol{\beta}$ is contained in a function of the data of the same dimensionality as $\boldsymbol{\beta}$.

[2] From the well known relations $E(\frac{\partial \mathcal{L}}{\partial \theta}) = 0$, $E(\frac{\partial^2 \mathcal{L}}{\partial \theta^2}) + E(\frac{\partial \mathcal{L}}{\partial \theta})^2 = 0$, and from (3.13), we have $\frac{\partial \mathcal{L}}{\partial \theta} = \frac{Y - b'(\theta)}{a(\psi)}$ and $\frac{\partial^2 \mathcal{L}}{\partial \theta^2} = -b''(\theta)/a(\phi)$

3.2.1 Poisson distribution

The Poisson probability distribution function is

$$f_i(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} \tag{3.11}$$

for $y_i = 0, 1, 2 \ldots$. The moments are $E(Y_i) = var(Y_i) = \mu_i$. We now verify that this distribution belongs to the exponential family. After taking log:

$$\log f_i(y_i) = y_i \log \mu_i - \mu_i - \log(y_i!)$$

we can easily recognize the canonical parameter in the coefficient of y_i

$$\theta_i = \log \mu_i,$$

so the canonical link is the logarithm. The second term in the pdf can be written as

$$b(\theta_i) = \exp(\theta_i),$$

and the last term is a function of y_i so we can identify

$$c(y_i, \phi) = -log(y_i!).$$

For simplicity, we can take $a_i(\phi) = \phi$ and $\phi = 1$. If we differenciate $b(\theta_i)$ we can verify the mean and variance:

$$\mu_i = b'(\theta_i) = \exp(\theta_i) = \mu_i,$$

$$\upsilon_i = a_i(\phi) b''(\theta_i) = \exp(\theta_i) = \mu_i.$$

3.3 Maximum Likelihood Estimation

Likelihood theory

Let $Y_1, \ldots, Y_n$ be n independent random variables with probability density functions (pdf) $f_i(y_i; \theta)$ depending on a vector-valued parameter θ. The joint density of n independent observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ is

$$f(\boldsymbol{y}; \theta) = \prod_{i=1}^{n} f_i(y_i; \theta) = L(\theta; \boldsymbol{y}). \tag{3.12}$$

Often is easier to work with natural logarithm of the likelihood function. A sensible way to estimate the parameter θ given the data $\boldsymbol{y}$ is to maximize the likelihood (or equivalently the log-likelihood) function, choosing the parameter value that makes the data actually observed as likely as possible.

This expression, viewed as a function of the unknown parameter θ given the data $\boldsymbol{y}$, is called the *likelihood function*.

Consider independent random variables $Y_1, \ldots, Y_N$ satisfying the properties of a generalized linear model. We wish to estimate parameters β which are related to the Y_i's through $E(Y_i) = \mu_i$ and $g(\mu_i) = \boldsymbol{x_i\beta}$. The log-likelihood function is

$$\mathcal{L} = \sum_{i=1}^{n} \mathcal{L}_i = \frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi) \tag{3.13}$$

To obtain the maximum likelihood estimator for the parameter β_j we need

$$\frac{\partial \mathcal{L}}{\partial \beta_j} = U_j = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \mathcal{L}_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_i} \tag{3.14}$$

using the chain rule for differentiation.We will consider each term on the right hand side of (3.14) separately. First, by differentiating((3.13)) and substituting ((3.9)):

$$\frac{\partial \mathcal{L}_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a_i(\phi)} = \frac{y_i - \mu_i}{a_i(\phi)}.$$

Next

$$\frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{\frac{\partial \mu_i}{\partial \theta_i}}.$$

Differentiation of (3.9) gives

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = var(Y_i)/a_i(\phi).$$

Finally, from (3.10)

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}. \tag{3.15}$$

Hence the score, given in (3.14), is

$$U_j = \sum_{i=1}^{n} \frac{y_i - \mu_i}{var(Y_i)} x_{ij} \frac{\partial \mu_i}{\partial \eta_i}. \tag{3.16}$$

The variance-covariance matrix of the U_j's has terms

$$\Im = E[U_j U_k]$$

which form the *information matrix* $\Im$. From (3.16)

$$\begin{aligned} \Im_{jk} &= E\left\{ \sum_{i=1}^{n} \left[\frac{Y_i - \mu_i}{var(Y_i)} x_{ij} \left(\frac{\partial \mu_i}{\partial \eta_i} \right) \right] \sum_{l=1}^{n} \left[\frac{Y_l - \mu_l}{var(Y_l)} x_{lj} \left(\frac{\partial \mu_l}{\partial \eta_l} \right) \right] \right\} \\ &= \sum_{i=1}^{n} \frac{E[(Y_i - \mu_i)]^2 x_{ij} x_{ik}}{[var(Y_i)]^2} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2 \end{aligned} \tag{3.17}$$

because $E[(Y_i - \mu_i)(Y_l - \mu_l)] = 0$ for $i \neq l$ as the Y_i's are independent. Using $E[(Y_i - \mu_i)^2] = var(Y_i)$, (3.17) can be simplified to

$$\Im_{jk} = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{[var(Y_i)]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2. \tag{3.18}$$

To find $\boldsymbol{b}^{(m)}$, the vector of estimates of the parameters $\beta_1, \ldots, \beta_p$, we can use the method of scoring which is, at the mth iteration

$$\boldsymbol{b}^{(m)} = \boldsymbol{b}^{(m-1)} + [\Im^{(m-1)}]^{-1}\boldsymbol{U}^{(m-1)}. \tag{3.19}$$

$[\Im^{(m-1)}]^{-1}$ is the inverse of the information matrix with elements $\Im_{jk}$ given by (3.18) and $\boldsymbol{U}^{(m-1)}$ is the vector of elements given by (3.16), all evaluated at $\boldsymbol{b}^{(m-1)}$. Multiplying both sides of (3.19) by $\Im^{(m-1)}$ we obtain

$$\Im^{(m-1)}\boldsymbol{b}^{(m)} = \Im^{(m-1)}\boldsymbol{b}^{(m-1)} + \boldsymbol{U}^{(m-1)}. \tag{3.20}$$

From (3.18) $\Im$ can be written as

$$\Im = \boldsymbol{X'WX}$$

where (W) is the nxn diagonal matrix with elements

$$w_{ij} = \frac{1}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2. \tag{3.21}$$

From equation (3.18) and (3.16), the expression on the right-hand side of (3.20) is the vector with elements

$$\sum_{k=1}^{p}\sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 b_k^{(m-1)} + \sum_{i=1}^{n}\frac{(y_i-\mu_i)x_{ij}}{var(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right) \tag{3.22}$$

evaluated at $\boldsymbol{b}^{(m-1)}$. Hence the right-hand side of (3.20) can be written as

$$\Im = \boldsymbol{X'Wz}$$

where $\boldsymbol{z}$ has elements

$$z_i = \sum_{k=1}^{p} x_{ik}b_k^{(m-1)} + (y_i - \mu_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \tag{3.23}$$

with μ_i and $\partial\eta_i/\partial\mu_i$ evaluated at $\boldsymbol{b}^{(m-1)}$. Thus the iterative equation (3.20), can be written as

$$\boldsymbol{X'WXb}^{(m)} = \boldsymbol{X'Wz}. \tag{3.24}$$

This is the same form as the normal equations for a linear model obtained by weighted least squares, except that it has to be solved iteratively because, in general, $\boldsymbol{z}$ and $\boldsymbol{W}$ depend on $\boldsymbol{b}$. Thus for generalized linear models, maximum likelihood estimators are obtained by an *iterative weighted least squares* procedure.

3.4 Likelihood Ratio Tests and the Deviance

The likelihood ratio criterion is a simple way for comparing any two nested models; now we will see how it can be constructed in this context [Dob08]. When we fit models to observations, the simplest possibility is the *null model*, which has only one parameter, thus all the variation are due to the random component. At the other extreme the *full model*, with one parameter per observation, assign all the cvariation to the systematic component.

As a matter of facts these are extreme models: the first being too simple and the second too uninformative, but they give us a guideline to understand the goodness of fit.

Consider first comparing a model of interest w with a *full model* Ω that provides a separate parameter for each observation. Let $\hat{\mu}_i$ denote the fitted values under w and let $\hat{\theta}_i$ denote the corresponding estimates of the canonical parameters. Similarly, let $\tilde{\mu}_0 = y_i$ and $\hat{\theta}_i$ denote the corresponding estimates under Ω.

The likelihood ratio criterion to compare two models in the exponential family has the form

$$-2\log\lambda = \frac{D(\boldsymbol{y}, \hat{\boldsymbol{\mu}})}{\phi}, \tag{3.25}$$

where the numerator, which does not depend on unknown parameters, is called the *deviance*:

$$D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = 2\sum_{i=1}^{n} p_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]. \tag{3.26}$$

We have use the fact that $a_i(\phi) = \phi/p_i$. The likelihood ratio criterion $-2logL$ is the deviance divided by the scale parameter ϕ, and is called the *scaled deviance*. For the Normal distribution the deviance is identical to the residual sum of squares, and minimum deviance is synonymous with least squares. If we need to compare two nested models w_1, with p_1 parameters, and w_2, with $p_2 > p_1$ parameters, the ratio of the maximized likelihoods can be written as a difference of deviance, since the maximizes log-likelihood under the saturated model cancels out. Hence we have

$$-2\log\lambda = \frac{D(w_1) - D(w_2)}{\phi}. \tag{3.27}$$

Large sample theory tells us that the asymptotic distribution of this criterion under the usual regularity conditions is χ^2_ν with $\nu = p_2 - p_1$ degrees of freedom.

3.4.1 Poisson deviance

Let $\hat{\mu}_i$ denote the m.l.e.of μ_i under the model of interest and $\tilde{\mu}_i = y_i$ denote m.l.e. under full model. The deviance is

$$\begin{aligned} D &= 2\sum[y_i \log y_i - y_i - \log y_i! - y_i \log\hat{\mu}_i + \hat{\mu}_i + \log y_i!] \\ &= 2\sum[y_i \log\frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)]. \end{aligned} \tag{3.28}$$

The Poisson deviance has an asymptotic chi-squared distribution as $n \to \infty$ with the number of parameters p remaining fixed, and can be used as a goodness of fit test. Differences between Poisson deviances for nested models (i.e. the log of the likelihood ratio test criterion) have asymptotic chi-squared distributions under the usual regularity conditions

3.5 GLM Multinomial

Multinomial distribution

Consider a set of n trials where

$$X_i = \begin{cases} 1, & i^{th} \text{ trial is a success} \\ 0, & n \text{ otherwise} \end{cases}$$

if $X_1, \ldots, X_n$ are indipendent with $P(X_i = 1) = p$, then $Y = \sum_{i=1}^{n} X_i$ has a $Binomial(n, p)$ distribution:

$$P(Y = y) = \binom{n}{y} p^y (1-p)^{n-y}, \quad y = 0, \ldots, n, \quad 0 < p < 1. \tag{3.29}$$

We have $E[Y] = np$ and $Var[Y] = np(1-p)$.

The multinomial distribution is a generalization of the binomial distribution where there are more than two possible response categories. Consider a set of n trials with d possible response categories. Let $\boldsymbol{X_i}$ be the d-dimensional response vector for trial i, where $X_{ij} = 1$ if response category j occurs on trial i and $X_{ij} = 0$ otherwise. Let $\boldsymbol{X_1, \ldots, X_n}$ be indipendent with $P(X_{ij} = 1 = \theta_j$; let $\boldsymbol{Y} = \sum_{i=1}^{n} \boldsymbol{X_i}$. Then Y_j is the number of response in category j. Then $\boldsymbol{Y}$ has a multinomial distribution:

$$P(\boldsymbol{Y = y}) = \frac{n!}{y_1! \ldots y_d!} \theta_1^{y_1} \ldots \theta_d^{y_d}, \tag{3.30}$$

$0 \leq y_1, \ldots, y_d \leq n, \quad \sum_{j=1}^{d} y_i = n, \quad 0 \leq \theta_1, \ldots, \theta_d < 1$. We have $E[Y_j] = n\theta_j$ and $Var[Y_j] = n\theta_j(1-\theta_j)$.

Multinomial distribution is connected with Poisson distribution. Consider indipendent r.v.'s $Y_1 \sim Poisson(\mu_i)$ and $Y_2 \sim Poisson(\mu_2)$. If we know that $Y_1 + Y_2 = n$, then

$$\begin{aligned} P(Y_1 = y_1, Y_2 = y_2 | Y_1 + Y_2 = n) &= \frac{P(Y_1 = y_1, Y_2 = n - y_1)}{P(Y_1 + Y_2 = n)} \\ &= \frac{e^{-\mu_1}\mu_1^{y_1}}{y_1!} \frac{e^{-\mu_2}\mu_2^{n-y_1}}{(n-y_1)!} \frac{1}{\frac{e^{-\mu_1-\mu_2}(\mu_1+\mu_2)^n}{n!}} \\ &= \binom{n}{y_1} \left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{y_1} \left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{n-y_1} \end{aligned} \tag{3.31}$$

In general, if we have d counts, and again condition on their sum, we will end up with the multinomial distribution,

$$P(Y_1 = y_1, \ldots, Y_d = y_d | \sum_{i=1}^{d} Y_i = n) = \frac{n!}{y_1! \ldots y_d!} \left(\frac{\mu_1}{\sum_{i=1}^{d} \mu_i}\right)^{y_1} \ldots \left(\frac{\mu_d}{\sum_{i=1}^{d} \mu_i}\right)^{y_d} \tag{3.32}$$

We cannot use GLM to model a multinomial response since the multinomial distribution is not in the 1-parameter exponential family. However, it turns out that we can estimate the parameters in the model using a Poisson GLM with log link, if we choose the proper linear predictor: the MLEs of multinomial logit models can be obtained by fitting a Poisson GLM with log link as long as the parameters that correspond to the fixed marginal totals are included in the model in the appropriate way [Agr13]. We have already seen that if $Y_1, Y_2, ..., Y_n$ are independent Poisson random variables, then the joint distribution of $Y_1, Y_2, ..., Y_n$ conditional on the total count $\sum_{i=1}^{n} Y_i$ is multinomial. In other words, the likelihood associated with

multinomial observations is the same as the likelihood associated with Poisson observations that are constrained by a fixed total.

Let $Y_i \sim Poisson(\mu_i)$, where

$$\log \mu_i = \phi + \boldsymbol{x_i}\boldsymbol{\beta}.$$

The log-likelihood is then (we do not write the term $-\log y_i!$ because it do not depend on β):

$$\begin{aligned} \log \mathcal{L}(\beta) &= \sum_i y_i \log \mu_i - \sum_i \mu_i \\ &= \sum_i y_i(\phi + \boldsymbol{x_i}\boldsymbol{\beta}) - \sum_i \mu_i \\ &= \phi \sum_i y_i + \sum_i y_i \boldsymbol{x_i}\boldsymbol{\beta} - \sum_i \mu_i. \end{aligned} \tag{3.33}$$

Let $m = \sum_i y_i$ and let $\tau = \sum_i \mu_i$. Then we can derive an expression for ϕ in terms of τ:

$$\tau = \sum_i \exp \phi + \boldsymbol{x_i}\boldsymbol{\beta} = \exp \phi \sum_i \exp \boldsymbol{x_i}\boldsymbol{\beta} \tag{3.34}$$

$$\phi = \log \tau - \log \sum_i \exp \boldsymbol{x_i}\boldsymbol{\beta}. \tag{3.35}$$

Now we can reparameterize the log-likelihood in terms of τ and β:

$$\begin{aligned} \log \mathcal{L}(\tau, \beta) &= m\left[\log \tau - \log \sum_i \exp \boldsymbol{x_i}\boldsymbol{\beta}\right] + \sum_i y_i \boldsymbol{x_i}\boldsymbol{\beta} - \tau \\ &= (m \log \tau - \tau) + \left[\sum_i y_i \boldsymbol{x_i}\boldsymbol{\beta} - m \log \sum_i \exp\{\boldsymbol{x_i}\boldsymbol{\beta}\}\right] \end{aligned} \tag{3.36}$$

The first term is the log-likelihood associated with $\sum_i Y_i$ (recall that we are assuming that we have observed $\sum_i y_i = m$ and that $\sum_i Y_i \sim Poisson(\sum_i \mu_i)$ where $\sum_i \mu_i = \tau$.) The second term is the log-likelihood associated with $Y_1, Y_2, \ldots, Y_n$ conditional on $\sum_{i=1}^n Y_i = m$, i.e. the multinomial logit model! To see this:

$$\begin{aligned} \log P\left(Y_1 = y_1, \ldots, Y_n = y_n | \sum_{i=1}^n Y_i = m\right) &= \sum_{i=1}^n y_i \log\left(\frac{\mu_i}{\sum_{j=1}^n \mu_j}\right) \\ &= \sum_{i=1}^n y_i \log\left(\frac{\exp\{\phi + \boldsymbol{x_i}\boldsymbol{\beta}\}}{\sum_{j=1}^n \exp\{\phi + \boldsymbol{x_j}\boldsymbol{\beta}\}}\right) \\ &= \sum_{i=1}^n y_i \log\left(\frac{\exp\{\boldsymbol{x_i}\boldsymbol{\beta}\}}{\sum_{j=1}^n \exp\{\boldsymbol{x_j}\boldsymbol{\beta}\}}\right) \\ &= \sum_{i=1}^n y_i \boldsymbol{x_i}\boldsymbol{\beta} - m \log\left(\sum_{j=1}^n \exp\{\boldsymbol{x_j}\boldsymbol{\beta}\}\right) \end{aligned} \tag{3.37}$$

We notice that all of the information about the parameter of interest, β, resides in the second term. The implication is that the MLE of β and its approximated asymptotic variance is the same regardless of whether we use the full log-likelihood $\log \mathcal{L}(\tau, \beta)$ (corresponding to the Poisson model) or just the log-likelihood for the multinomial logit model! The only issue is that, when using the Poisson GLM to estimate β when the counts are actually multinomial, we have to incorporate nuisance parameter(s) (e.g., τ) in the model in the appropriate way.

3.6 GLM on USA commuting flows

Now we apply the technique that we have descibed in this chapter to fit data of migration flows of an area of the USA with 300 counties (see chapter 6 for description of dataset) with different kinds of Gravity Models.

We start with

$$\log \mu_{ij} = \log k + \alpha \log P_i + \beta \log P_j - \gamma \log d_{ij}, \tag{3.38}$$

fitted with ordinary least squares (OLS). In table (3.1) we show the results. We notice that the sum of the estimated flows is considerably less than the sum of the observed flows. Clearly it is not a good fit.

Then we apply GLM with poisson distribution and logarithm as the link function, using equations (3.4) and (3.38). To compare this model with the previous one we use the chi-square, which reveals that the second is better, as expected. In this case the sum of observed flows is very close to the sum of estimated flows (table (3.1)) .

Table 3.1: Results of fit of the Gravity Model 3.38 with OLS and GLM. We report parameters ($\log(k)$, α, β, γ), chi-square, and sum of observed and estimated flows.

OLS	parameter $[2.83 \pm 0.05$, 0.218 ± 0.003 , 0.301 ± 0.003 , $1.542 \pm 0.006]$
	chi-square $8.99e+08$, sum observed flows 5209095 , sum estimated flows 188517
GLM	$[-2.375 \pm 0.005, 0.583 \pm 0.001, 0.8042 \pm 0.0001, 2.0249 \pm 0.0001]$
	chi-square $1.05e+07$, sum observed flows 5209095, sum estimated flows 5208408

The effect of distance on trip-making behaviour is inverse to the power λ. However, as the distance between i and j becomes smaller and smaller (approaching zero), the number of trips the model predicts becomes larger and larger (approaching infinity). To avoid this problem, a better function of distance would be the negative exponential.

So we fit with a negative exponential as the deterrence function, using (3.4) with

$$\log \mu_{ij} = \log k + \alpha \log P_i + \beta \log P_j - \gamma d_{ij}. \tag{3.39}$$

Looking at the results in table (3.2) and comparing the deviance of this model with the previous one (with power-law as the deterrence function) we find that exponential is a better deterrence function.

In the end we use another approach. We fit the model with a Multinomial distribution instead of Poisson; this means that we consider fixed the number of people who migrate from each location (singly-costrainted Gravity Model). As we show in section 3.5, GLM with Multinomial distribution is equivalent to GLM with Poisson distribution, with a vector of nuisance parameter, one for each location, which must be included in the model in order to satisfy the condition that we have fixed the number of migrants from every county. From this fit we find more parameters, but we are interested only in the usual ones. The formula used is

$$\log \mu_{ij} = k + \alpha \log P_i + \beta \log P_j - \gamma d_{ij} + \tau_i. \tag{3.40}$$

We fit with both negative exponential and inverse power-law as the deterrence function. Looking at the deviance in table (3.2), we confirm again that the negative exponential gives

a slightly better fit; we also check that nuisance parameters are in agreement with the expected total. If we compare the multinomial model with the corrisponding Poisson version (with the deviance, table (3.2)), we find that the first one provides the better fit. It could be argued that the multinomial model depends on more parameters, and this may be the reason for the better fit. If we subtract the deviances and we consider the difference of the degrees of freedom, as in the deviance criteria in section 3.4, we see that the multinomial model can explain the data better(deviance difference: 2088000.0, d.o.f. 298, p-value ~ 0).

Table 3.2: Results of fit of different Gravity Models:
poisson distribution of trips and power-law as the deterrence function (poisson pow),
poisson distribution of trips and negative exponential as the deterrence function (poisson exp),
multinomial distribution of trips and power-law as the deterrence function (multinomial pow),
multinomial distribution of trips and negative exponential as the deterrence function (multinomial exp).
Intervening opportunities models:
poisson distribution of trips and power-law as the deterrence function (io pow),
multinomial distribution of trips and power-law as the deterrence function (io mult pow),
We report parameters ($\log(k)$, α, β, γ) of every models (for multinomial we omit $\log(k)$), and the deviance. We can notice that the best model is the singly-costrainted Gravity Model with negative exponential as deterrence function.

model	parameters	deviance
poisson pow	$[-2.375 \pm 0.005,\ 0.583 \pm 0.001,\ 0.8042 \pm 0.0001,\ 2.0249 \pm 0.0001]$	$6.79e+06$
poisson exp	$[-4.449 \pm 0.005, 0.4637 \pm 0.0001, 0.7181 \pm 0.0001, 0.0521 \pm 0.0001]$	$5.02e+06$
multinomial pow	$[0.4935 \pm 0.001,\ 0.969 \pm 0.001\ ,\ 2.840 \pm 0.001]$	$3.72e+06$
multinomial exp	$[-0.0947 \pm 0.001,\ 0.9107 \pm 0.001,\ 0.0641 \pm 0.0001]$	$2.93e+06$
io pow	$[-4.0435 \pm 0.005,\ 1.5794 \pm 0.001\ ,\ 1.0050 \pm 0.0001,\ 1.4879 \pm 0.0001\ ]$	$6.66e+06$
io mult pow	$[1.2189 \pm 0.001,\ 1.0304 \pm 0.001,\ 1.5303 \pm 0.0001]$	$5.01e+06$

In the end we apply intervening opportunities model, using also in this case population as an estimate of attractiveness of counties. We assume for trip probability the expression:

$$p_{ij} = k\frac{P_i^{\alpha}P_j^{\beta}}{v_{ij}^{\gamma}}, \tag{3.41}$$

where now the distance at the denominator is replaced with v_{ij}, which reppresents the number of intervening opportunities between location i and j, i.e. the sum of the population of i and the population of the other cities in the middle. From results in table (3.2), we can see that with our choice of deterrence function, Gravity Models give better results, also in the case of Multinomial distribution.

We use also Probability Integral transform (PIT) to test the goodness of fits (more details about this technique in chapter 5). In figure 3.1, 3.2, 3.3 we show observed-estimated plot and the histogram of the PIT for

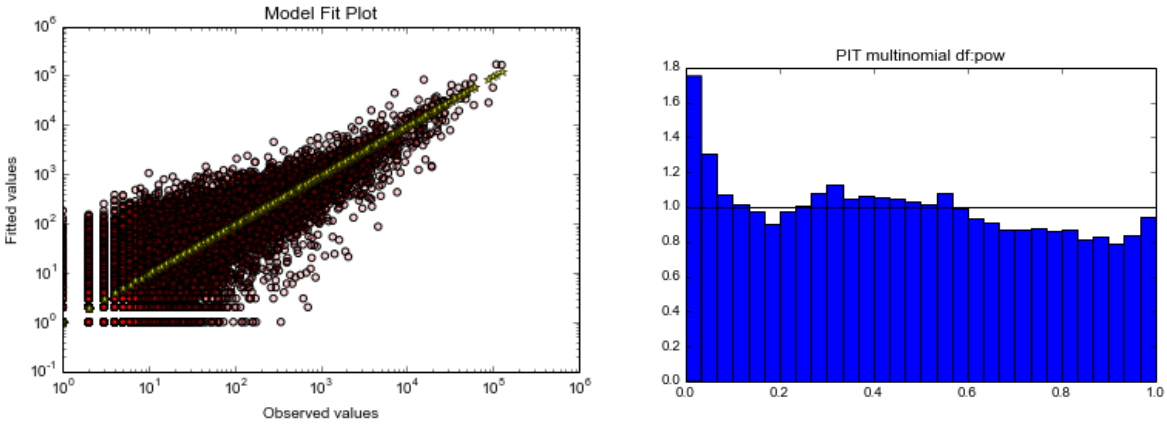

Figure 3.1: Results of the fit of Gravity Model with GLM; the probability distribution of trips is multinomial and the deterrence function is a power-law. On the left: observed-estimated flows; on the right: probability integral transform (distance from uniform distribution: MSE 0.031). The minor deficiences of the model can be seen in the difference between the probability integral transform and the uniform distribution.

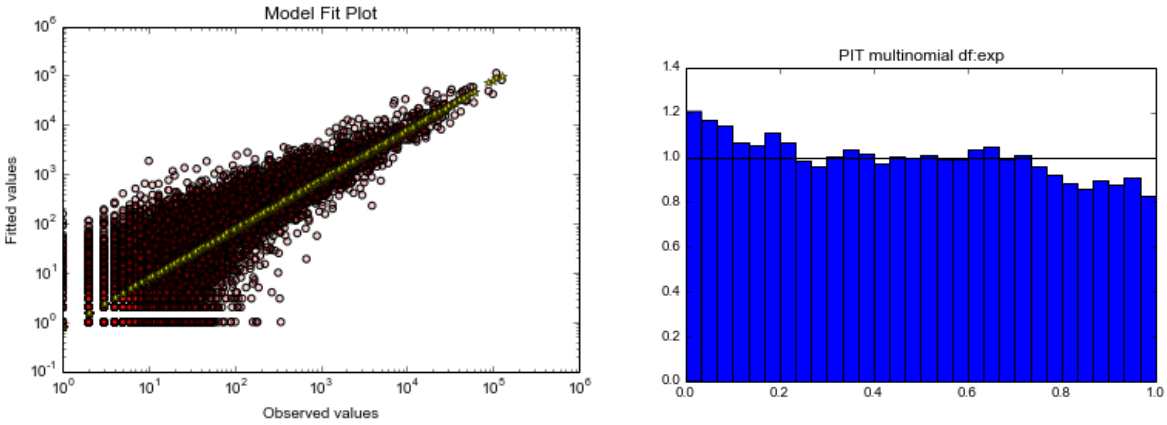

Figure 3.2: Results of the fit of Gravity Model with GLM; the probability distribution of trips is multinomial and the deterrence function is a negative exponential. On the left: observed-estimated flows; on the right: probability integral transform (distance from uniform distribution: MSE 0.007). The minor deficiences of the model can be seen in the difference between the probability integral transform and the uniform distribution.

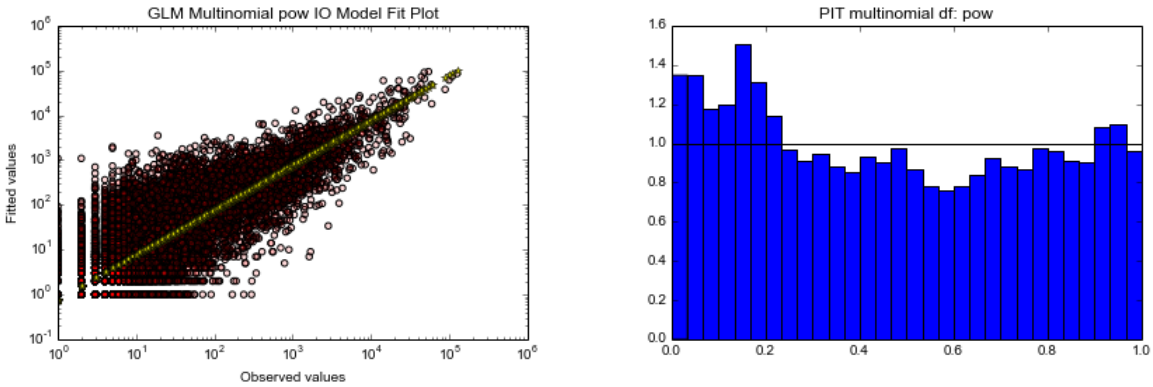

Figure 3.3: Results of the fit of Intervening opportunities model with GLM; the probability distribution of trips is multinomial and the deterrence function is a power-law. On the left: observed-estimated flows; on the right: probability integral transform (distance from uniform distribution: MSE 0.0338). The poor fit of the model can be deduced from the difference between the probability integral transform and the uniform distribution.

Chapter 4

Non-parametric method to estimate parameters

Here we propose a general method to find the deterrence function and the values of weights that provide an optimal estimate of the observed OD matrix, T_{ij} , using a *singly-constrained* gravity model. In singly-constrained gravity models the average number of trips from i to j is given by the following equation: $T_{ij} = T_i p_{ij}(w, f_\gamma)$, where:

$$p_{ij} = \frac{w_j f(r_{ij}, \gamma)}{\sum_k w_k f(r_{ik}, \gamma)} \tag{4.1}$$

is the estimated probability of a trip from location i to j, and depends on the weights, the deterrence function f_γ and its parameters.

Due to the independence of individual trips, the probability to observe the trips $\{T_{ij}\}$ from i to j , T_{ij}, is given by the multinomial distribution:

$$P(\{T_{ij}\}|\{p_{ij}\}, T_i) = T_i! \prod_k \frac{p_{ik}^{T_{ik}}}{T_{ik}!}. \tag{4.2}$$

As we see in chapter 2, usually the weight is assumed to be proportional to the resident population or the number of jobs, and in other cases it is assumed to be a function of these variables. The deterrence function is usually assumed to be power law, exponential, stretched exponential or a more complex function.

Our aim is to understand if the fundamental assumption of the singly-costrained Gravity Model (4.1) (4.2) are compatible with observed flows. In effect discrepancies between data and model can be due to two different kind of reasons:

1. fundamental assumptions (4.1) (4.2) are correct, but additional assumptions that usually are made on the particular deterrence function form or in the choice of weights are wrong.
2. fundamental assumptions (4.1) (4.2) are incorrect (for example, individual decision are not indipendent).

Our method should help us to decide if we are in the situation (1) or (2).

Our method will seek to find the singly-constrained gravity model that provides the best estimate of the observed flows. The difference of our approach with respect to other methods in the literature is that we do not make any a priori assumption on the functional

form of weights and deterrence function. In particular, weights do not explicitly depend on population or any other socio-economic variable, but are treated as free parameters. Similarly, the deterrence function is estimated using a non-parametric regression, i.e. without pre-imposing a particular functional form but expressing it as a sum of 10 gaussians whose parameters (amplitude, mean, and variance) are free parameters. Our method thus consists in looking for the values of the n weights and the parameters of the 10 gaussians (i.e. n + 30 total parameters) that minimize a cost function that describe distance between model and data trips.

At the beginning we perform some numerical experiment, in order to test the ability of our methods to correctly estimate the deterrence function and the weights. We choose the most useful approach comparing slightly different algorithm.

4.1 Minimization techniques

The problem consists in estimating the true values of the weights w_i and to recover the deterrence function f, given the observed trips T_{ij}. To this end we test different approaches. The first is based on *downhill simplex algorithm*. It consist on two steps:

1. *Optimizing the deterrence function*: keeping the weights fixed, the downhill simplex algorithm is used to find the parameters of the 10 gaussians $\{A_k, \mu_k, \sigma_k\}_{k=1}^{10}$ that minimize the cost function.
2. *Optimizing the weights*: keeping the deterrence function's parameters fixed, the downhill simplex algorithm is used to find the weights w that minimize the cost function.

Steps 1. and 2. are repeated until the final values of the parameters do not significantly vary with respect to the initial values (typically when they have not changed more than 1%).

As a cost function, $f_c(\{A_k, \mu_k, \sigma_k\}_{k=1}^{10}, w)$, first we use the absolute value of the difference of the estimate flows (T_{ij}) from the observed flows (T_{ij}^*), in order to find the model with the shortest L_1 distance from the data,

$$f_c = \sum_{ij} |T_{ij} - T_{ij}^*|. \tag{4.3}$$

We use also another cost function, derived from the likelihood. The likelihood of a set of parameter values, ($\theta = \{A_k, \mu_k, \sigma_k\}_{k=1}^{10}, w$), given outcomes T_{ij}, is equal to the probability of those observed outcomes given those parameter values. For a multinomial distribution we have (4.2):

$$\mathcal{L}(\theta|T_{ij}) = P(T_{ij}|\theta) = \prod_{ij} P(\{T_{ij}\}|T_i, \{p_{ij}\}) = \prod_{ij} T_i! \prod_k \frac{p_{ik}^{T_{ik}}}{T_{ik}!},$$

we take the logarithm:

$$\log \mathcal{L}(\theta|T_{ij}) = \sum_{ij} \log(T_{ij}!) + \sum_k T_{ik} \log\left(p_{ik}(\theta)\right) - \sum_k \log T_{ik}!. \tag{4.4}$$

We need to maximize the probabilities respect to the parameters θ, so we can consider only the term in 4.4 which depends on θ. The cost function we need to minimize simply

becomes:

$$f_c(\theta) = -\sum_k T_{ik} \log\left(p_{ik}(\theta)\right). \tag{4.5}$$

We use also a second approach based on *Greedy algorithm*. A greedy algorithm is a mathematical process that looks for simple, easy-to-implement solutions to complex, multi-step problems by deciding which next step will provide the most obvious benefit. Such algorithms are called greedy because while the optimal solution to each smaller instance will provide an immediate output, the algorithm does not consider the larger problem as a whole. In this context, we start from some initial value of the parameter θ. For each iteration we randomly choose one parameter and we change its value adding a normal variable with variance proportional to the parameter value. Then we check if the cost function has decreased after this change; only if this happen we accept the change. The process finish when the difference between the cost function in the last iteraction and the cost function of 100 iteration before is less than a small percent (0.001%).

With this method it is possible to impose condition on the parameter: in particular we need positive amplitude of gaussians and, especially, positive weights.

4.2 Numerical simulations

We perform two different kind of numerical simulation. In the first one (A), we consider a square of side $L = 500$, and we randomly place $n = 100$ points (locations) on it, extracting their x and y coordinated from a uniform distribution between 0 and 500. The deterrence function is assumed to be stretched exponential

$$f(r) = e^{-r^{1.5}/1500}; \tag{4.6}$$

we assigned a weight w to each locations extracting a number between 1,000 and 100,000 with uniform probability and the total number of trips departing from each location i, T_i, is randomly chosen between $0.5w_i$ and w_i. The matrix p_{ij} containing the probabilities of one trip between any two locations i and j is generated, and a realization of the OD matrix T_{ij} is extracted from the multinomial distribution 4.2.

The second simulation (B) is hybrid, because flows data are simulated started from real data of locations and the weights are assumed to be the real population of the city.

The method based on Downhill simplex algorithm works well with numerical simulation. We study sample (A) of synthetic data, and we use as initial values for the weights w_0 once the uniform value ($w_0 = 1$), and then total arrival for each locations. We use also both likelihood (4.4) and absolute value of difference (4.3) as cost function. In figure 4.1 the plot of observed-simulated value of flows for $w_0 = 1$ and cost function (4.3) are shown. The plot for the other combinations of initial values and cost functions look very similar. The only exception is $w0 = 1$ and likelihood as cost function: this combination does not converge to the right solution. If we calculate the distance (MSE) of observed-estimated value, we find that using likelihood leads to a distance slightly bigger. For example, MSE distance with likelihood is 1205, as with the other cost function (and arrival as initial weights) it is 444. In figure 4.2 and 4.3 the estimated deterrence function and the estimated weights are shown.

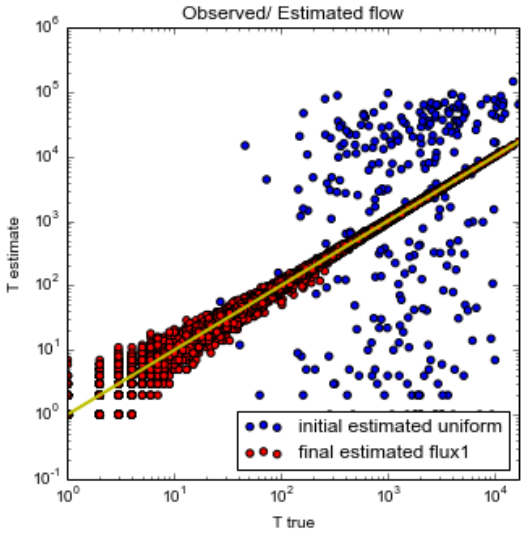

Figure 4.1: First algorithm applied on synthetic data (A). Plot of observed-simulated value of flows for $w_0 = 1$ and cost function (4.3) (absolute value of the difference between observed-estimated flows).

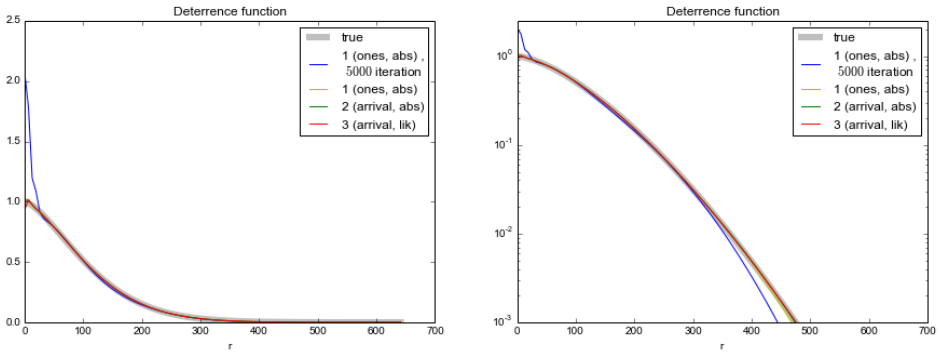

Figure 4.2: First algorithm applied on synthetic data (A). Deterrence function for the different combination of initial values and cost function. On the right side, logarithmic scale. We can also see an intermediate step for the deterrence function for the combination 1. The function is already very close to the true function.

Function are normalized in order to have the same integral (with this model both deterrence function and weights are defined apart from a normalization constant). We observe that the parameters of the deterrence function converge very quickly to their optimal values, well before the weights do. Moreover, values of the deterrence function's parameters close to the optimal ones are attained even when weights are far from their optimal (true) values. In this case, however, the estimated trips are very different from the observed (true) trips, i.e. the model's precision is mostly affected by the values of weights. It is thus important to carefully perform the weights optimization step.

At first we use as initial condition a sum of identical gaussian for the deterrence function, as for the weights we start from uniform values. With this algorithm the results depend on the initial values, and we see that the result is better if we use as initial weight for a location the total arrival in that place. The reason is that there is a strong correlation between total arrival and weights of a location; in figure (4.4) we can see results from synthetic data which follow multinomial distribution.

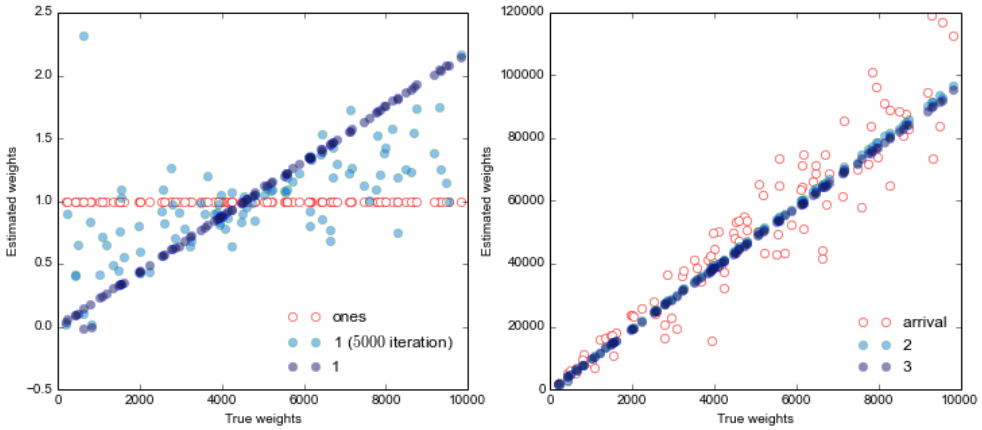

Figure 4.3: First algorithm (based on downhill simple algorithm) applied on synthetic data (A), plot of estimated-true weights with different initial values and cost function. On the left, initial values for weights $w_0 = 1$ and cost function (4.3) (absolute value of the difference between observed-estimated flows); we show initial values, an intermediate step (after 5000 iteration) and the final values. We can see that in the intermediate step, differently from the deterrence funcion, the weights are still far from their real values. On the right, initial values for weights w_0 =arrival and cost function in (2): (4.3) (absolute value of the difference between observed-estimated flows) and in (3) (4.4) (likelihood). We can see that we manage to predict well the true values of weights for every choice of the initial values or cost function.

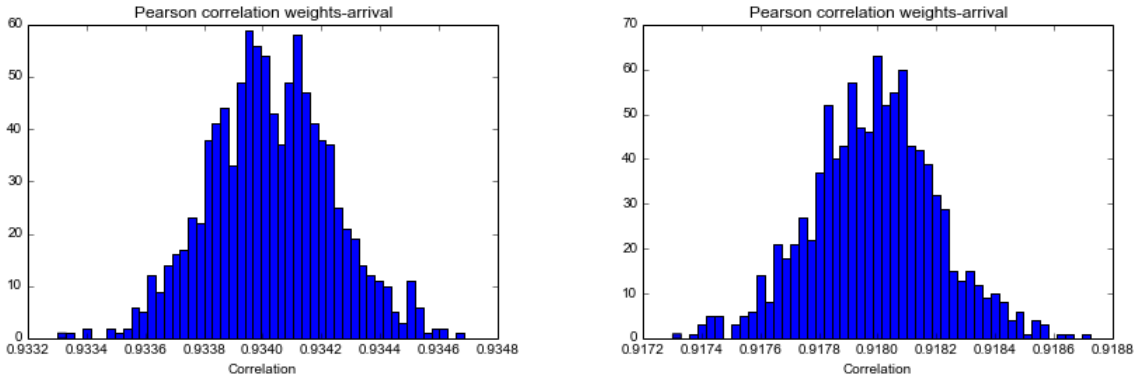

Figure 4.4: Correlation weights-arrival in 1000 hybrid simulation from data of migration in the USA (years 2000-2001), with population as weights. On the left, results with the deterrence function: stretched exponential (5.5). On the right, results obtained with the deterrence function: sum of gaussian (with parameter of the fit of group of 300 locations in data SOI 2000-2001.

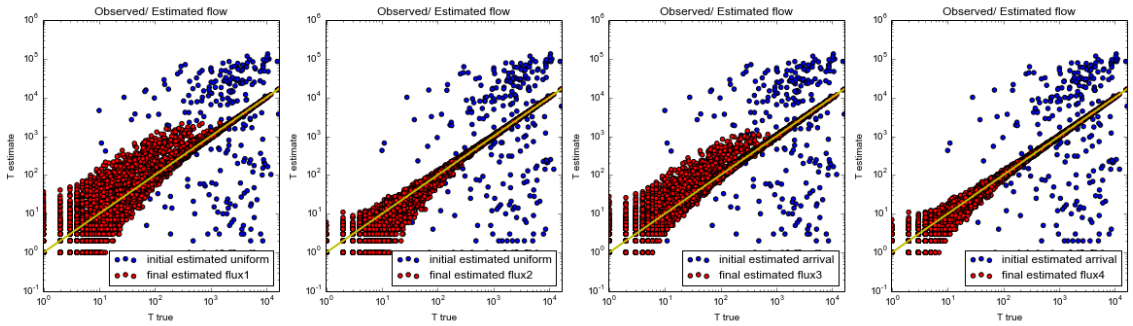

Figure 4.5: Greedy algorithm applied on synthetic data (A).
1:Initial weights: uniform values (one); initial parameter forcost function: absolute value of the difference between O/E flows (MSE 9329)
2:Initial weights: arrivals; initial parameter forcost function: absolute value of the difference between O/E flows (MSE 529).
3: Initial weights: uniform values (one); initial parameter forcost function: Likelihood (MSE 3650).
4: Initial weights: arrivals; initial parameter for cost function: Likelihood (MSE 470).
We can see that starting from arrivals as initial weights gives a better estimation of the real flows, as the two different choice for the cost function give equivalent results.

On synthetic data (A) we apply also Greedy algorithm. We start from different initial condition (uniform weights equal to one, or arrivals). The algorithm will stop after reach the same level of tollerance (0.0001%) for the final change in the cost function. We performed some simulation on the same combination of initial values-cost function of the simulation for the first algorithm; in this case we reached some level of convergence even with likelihood and uniform weights. In figure 4.6 we can see that the deterrence function converge always quite well to the real value. In figure 4.5 we see that if we start from uniform weights, the estimated flows are not exactly aligned with true flows as starting from the arrival. Looking for the MSE distance in the caption, we see that the performance of the two cost function are both quite good. This algorithm, as we expected, leads to different solution if we start from values of parameter too far from the real ones. In figure 4.8 we show the behaviour of the cost functon during the minimization process.

We then perform another simulation (B), with flows data closer to the real values, because we use real data of locations and the weights are assumed to be the real population of the city. The deterrence function is assumed to be stretched exponential (5.5). The simulation is performed for the whole USA, and after we select a region with 300 counties. In figure 4.9 we show the application of greedy algorithm on synthetic data (B). We use arrival as initial values for weights. For the initial values of the deterrence function we use a fit of $P(r)$, i.e. the probabilities density function of trip length, which is likely to be a function not so far from the deterrence function that we are looking for. The deterrence function derived from the fit is very close to the true function, and also the weights are very closed to the real ones. We always check the evolution of cost function (every 100 iteration), to be sure to have reach a minimum of the function.

To summarize, the first approach, based on downhill simplex algorithm, works well with synthetic data, but we prefere the Greedy algorithm because it gives us more control in the minimization process. In particular we can impose condition on the parameters.

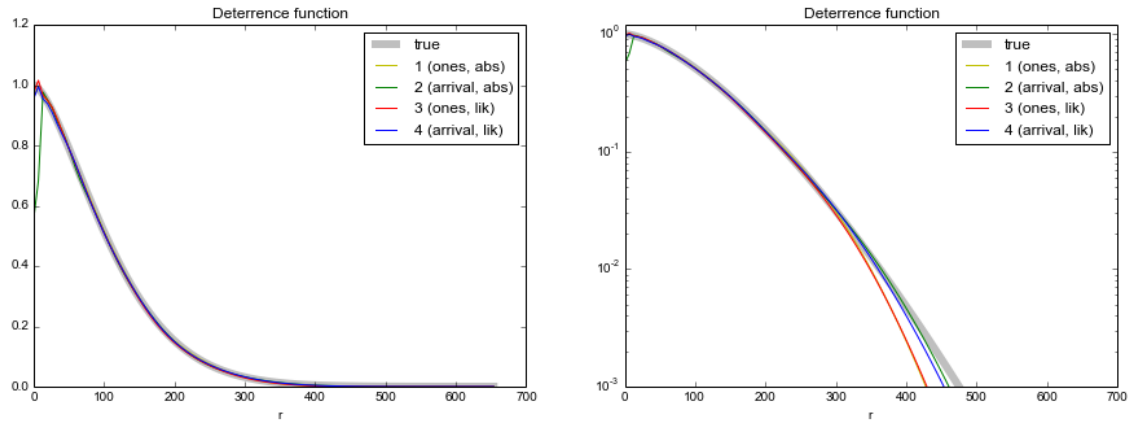

Figure 4.6: Deterrence function obtained with Greedy algorithm applied on synthetic data (A), with different combination of initial values and cost functions:
1:Initial weights: uniform values (one); initial parameter forcost function: absolute value of the difference between O/E flows.
2:Initial weights: arrivals; initial parameter forcost function: absolute value of the difference between O/E flows.
3: Initial weights: uniform values (one); initial parameter forcost function: Likelihood.
4: Initial weights: arrivals; initial parameter for cost function: Likelihood.
We can see that regardless of the initial values and the cost function, there is a perfect agreement with the true function.

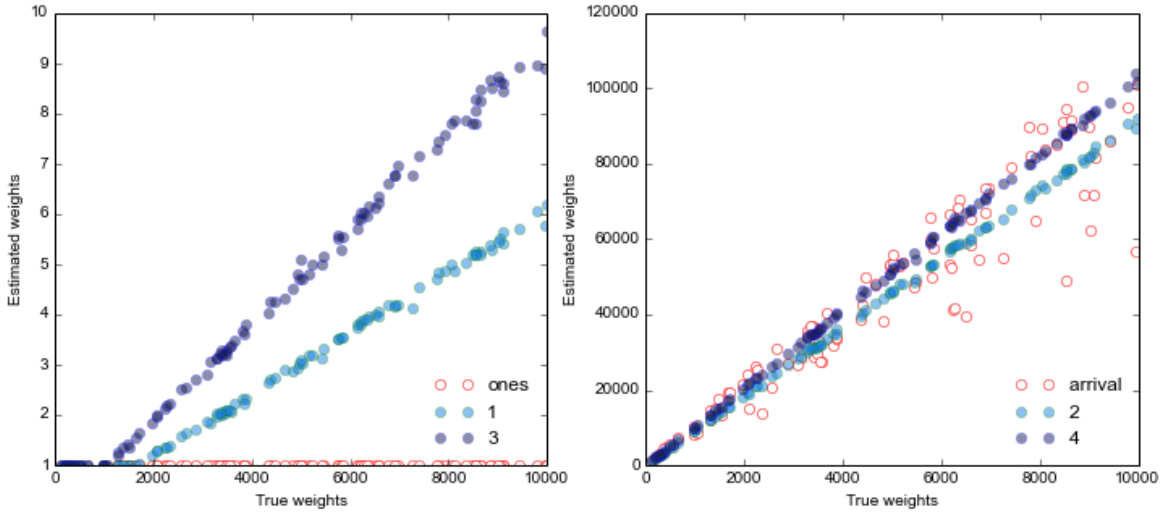

Figure 4.7: True-estimated weights for Greedy algorithm applied on synthetic data (A), with different combination of initial values and cost functions:
1:Initial weights: uniform values (one); initial parameter forcost function: absolute value of the difference between O/E flows.
2:Initial weights: arrivals; initial parameter forcost function: absolute value of the difference between O/E flows.
3: Initial weights: uniform values (one); initial parameter forcost function: Likelihood.
4: Initial weights: arrivals; initial parameter for cost function: Likelihood.
We notice that starting from arrivals gives better prediction of the true weights.

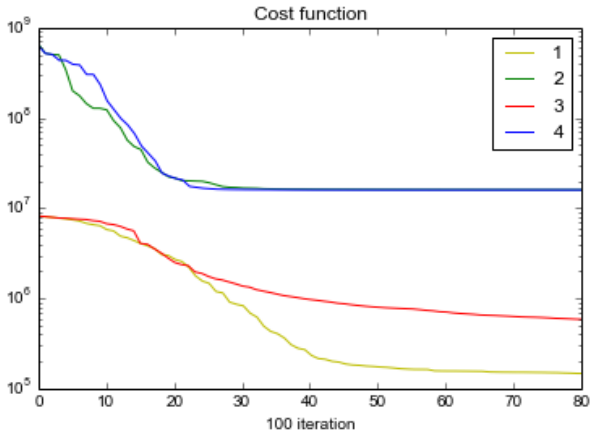

Figure 4.8: Cost function every 100 iteration of Greedy algorithm applied on synthetic data (A), with different combination of initial values and cost functions:
1:Initial weights: uniform values (one); initial parameter forcost function: absolute value of the difference between O/E flows.
2:Initial weights: arrivals; initial parameter forcost function: absolute value of the difference between O/E flows.
3: Initial weights: uniform values (one); initial parameter forcost function: Likelihood.
4: Initial weights: arrivals; initial parameter for cost function: Likelihood.
We notice that after a rapid initial decline (until ~ 4000 iterations), the function stabilises.

4.3 Aggregation techniques

We would like to apply this method to study spatial flows in the USA. The sample in this situation is very large (number of counties in the USA: 3144). Our method work easily with sample of around $100-400$ locations. So we decide to focus on a subset of USA counties, and in the meanwhile to study some technique of aggregation of locations. We try two methods:

- We divided locations using a fixed spatial *grid* (and so different number of counties in every aggregation). For every aggregation we consider as population the sum of all population. For the position we choose the midpoint of the coordinates of the counties. We consider only flows towards other counties outside the aggregation. We choose grid step in order to have set of $100-400$ locations to study (step: 170, 200, 250, 350 km).
- We aggregate locations in a *fixed group of city* (8, 10, 12, 14). We start aggregation from the counties close to the border and then we procede the selection towards the centre 4.10.

We perform some simulation aggregating synthetic data (B). In figure 4.11 we show some results of grid aggregation on these data. For grid $170km$ there is a good estimation of flows, and optimal weights are very close to population. Population is the real weight of our dataset; after aggregation, the actual optimal weights for the model may not necessarily be the sum of population of counties in the subset. In table 4.1 we report the correlation for the different grid steps. As expected, the correlation weights-population is very strong for grid with smaller step, and slowly it decrease for larger step.

Also the deterrence function present some deviation from the real one. We see that step $170km$ is for every aspect a good approximation, but larger steps are not satisfactory.

We test also the aggregation with fixed number of counties for synthetic data (B). In figure 4.12 we show some results. For aggregation of 8 number there is a good estimation of flows,

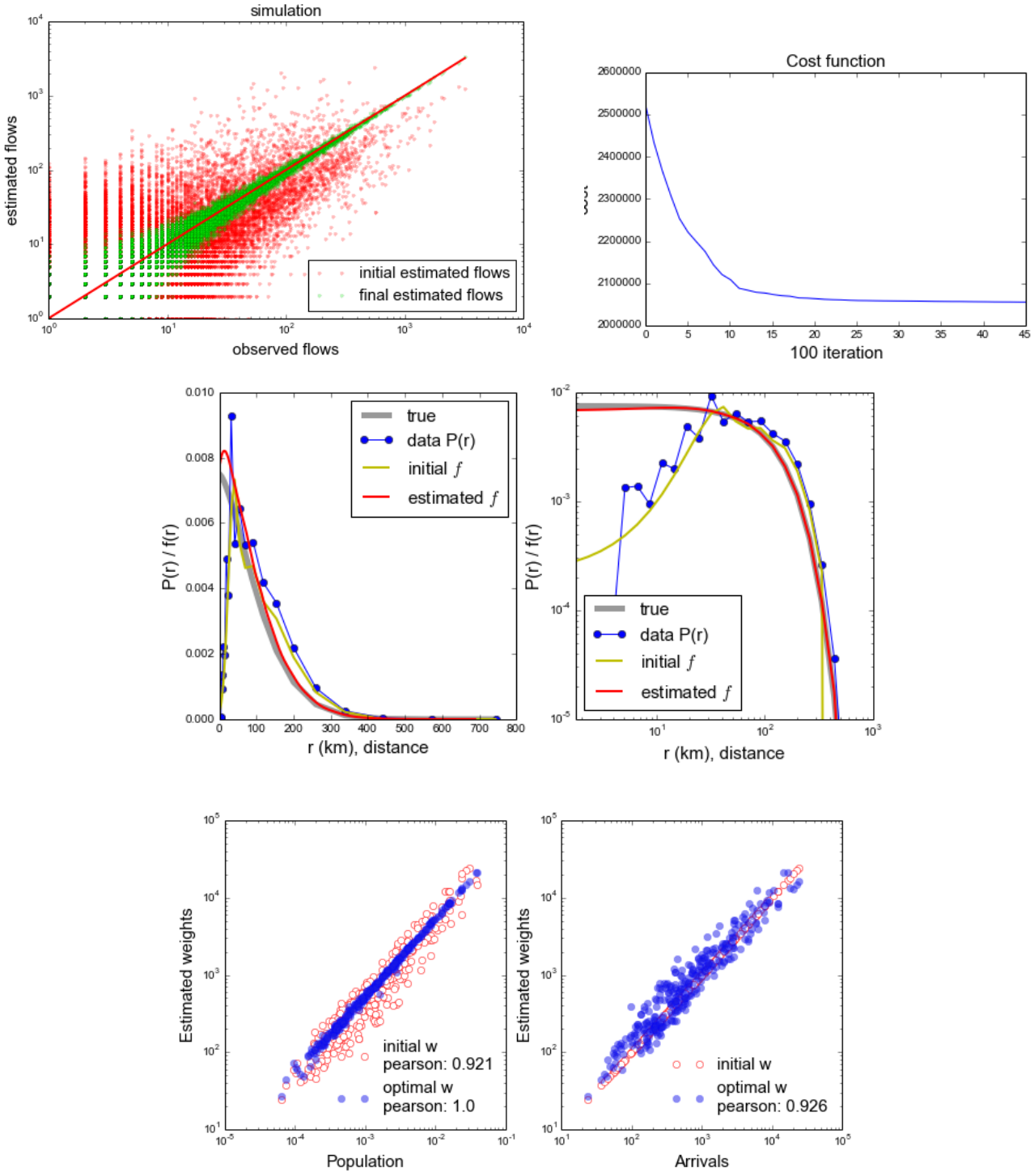

Figure 4.9: Simulation on data (B, simulated with real values of populations as weights and distances, from data migration in the USA (years 2000-2001), in a region of with 300 counties. Observed estimated flows, deterrence function (almost the same of the real one). Correlation between arrival and weights. Cost function every 100 iteration. The deterrence function derived from the fit is very close to the true function, and also the weights are very closed to the real ones. From the evolution of cost function we can check to have reached a mimimum of the cost function.

Table 4.1: Correlation values for aggregation in grid (of step: grid size) of simulated data. Total is the total number of locations considered after aggregation. We report the correlation between population and arrivals, weights and arrivals, weights and population. Population correspond to true value of weights before aggregation; the correlation weights-population is very strong for grid with smaller step, and slowly it decrease for larger step.

grid size	total	population/arrivals	weights/arrivals	weights/population
170 km	321	0.891	0.887	0.981
200 km	240	0.902	0.91	0.935
250 km	155	0.905	0.912	0.922
350 km	86	0.91	0.892	0.882

and optimal weights are very close to population. The estimation of the deterrence function is quite good for every different aggregation, but from the correlation weights-population in table (4.1) we can see that the the weights go very quickly far from the population as the number of aggregated counties increase.

Table 4.2: Correlation values for aggregation in fixed number of counties (# counties) of simulated data. Total is the total number of locations considered after aggregation. We report the correlation between population and arrivals, weights and arrivals, weights and population. Population correspond to true value of weights before aggregation; from the correlation weights-population we can notice that weights go very quickly far from the population as the number of aggregated counties increase.

# counties	total	population/arrivals	weights/arrivals	weights/population
8	389	0.746	0.774	0.981
10	311	0.725	0.807	0.964
12	259	0.837	0.472	0.577
14	222	0.888	0.41	0.505

So we conclude that with synthetic data both this two kind of aggregation are good approximation if we choose the finest aggregation (step $170km$ or 8 number of cities). We notice that the deterrence function that we choose for generating synthetic data has a range of about 300 km, and, even after aggregation, average counties dimension ($\sim$ 100 km) is shorter than the range.

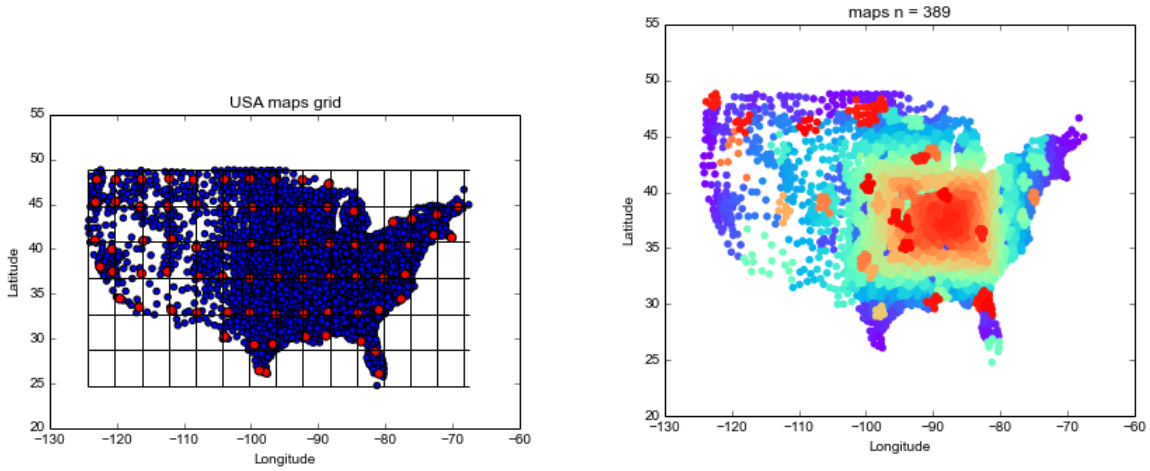

Figure 4.10: On the left, example of aggregation of counties in the USA with a grid of step $400km$. On the right, aggregation in group of 8 cities.

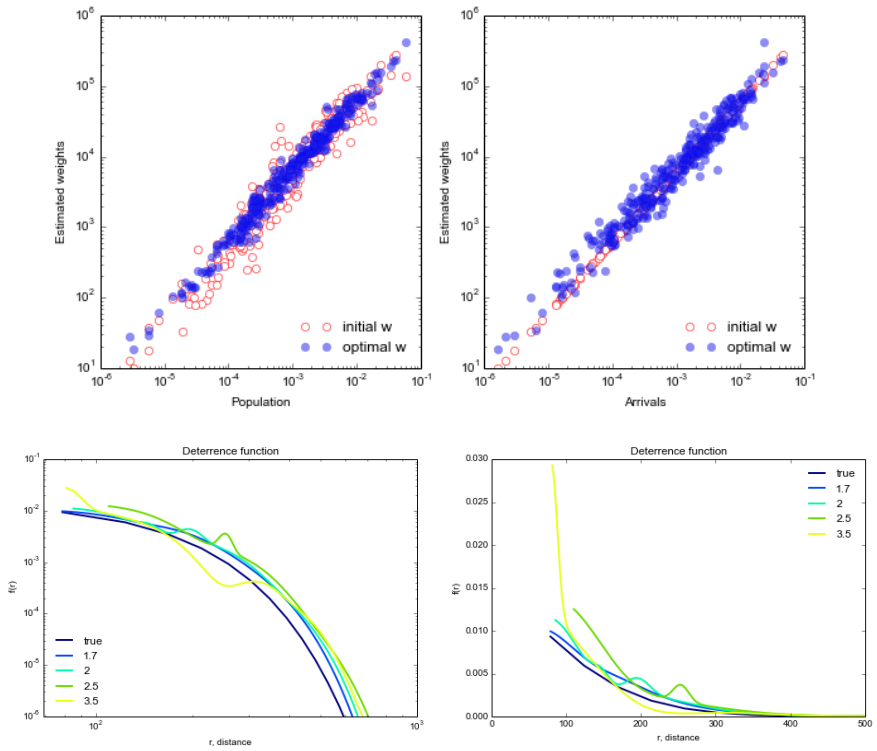

Figure 4.11: Grid aggregation on synthetic data (B). Estimated weights-population and estimated weights-arrival for grid of step 170km. Comparison between deterrence function (log - linear scale) of different grid step.

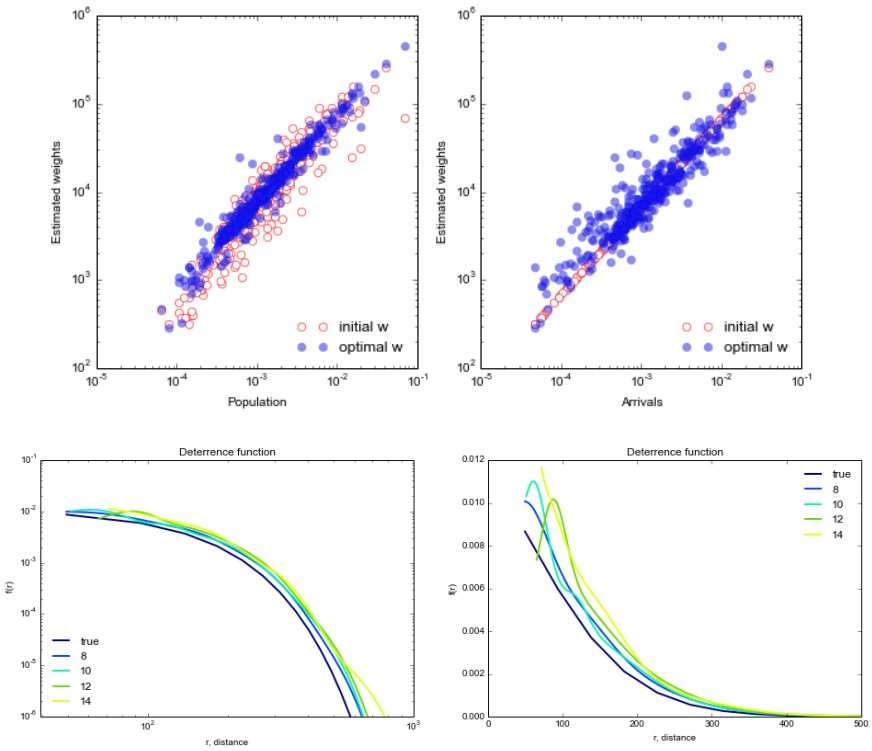

Figure 4.12: Aggregation with fixed number of cities on synthetic data (B). Estimated weights-population and estimated weights-arrival for aggregation of 8 cities. Comparison between deterrence function (log - linear scale) of different aggregation number.

Chapter 5

Test of spatial model assumptions

Spatial models of human mobility and interactions, like the Gravity Model and the Intervening Opportunities model, rely on some common assumptions and differ on the variable they use to predict the probability of one flow event.

The common assumptions are:

- Individual events are independent. For example, the decision of individual A to commute to location j is independent of the decision of individual B (irrespective of the distance of the home locations of A and B).
- The probability p of a flow event from location i to location j depends on
  - one variable per location, w_i (e.g. population within some distance, number of calls, ...).
  - a variable relating two locations, r_{ij} (e.g. geographic distance for GM, intevening opportunities for IOM).

Then p has the following form

$$p_{ij} = \frac{w_j f(r_{ij})}{\sum_k w_k f(r_{ik})}. \tag{5.1}$$

It follows that the probability of observing the trip $\{T_{ij}\}$ from i is given by the multinomial distribution

$$P(\{T_{ij}\}|\{p_{ij}\}, T_i) = T_i! \prod_k \frac{p_{ik}^{T_{ik}}}{T_{ik}!} \tag{5.2}$$

and the average and variance of the number of the flows from i to j is respectively

$$\langle T_{ij} \rangle = T_i p_{ij} \tag{5.3}$$

and

$$\sigma_{ij}^2 = T_i p_{ij}(1 - p_{ij}). \tag{5.4}$$

Our aim is to understand if the common assumptions that we highlighted are compatible with a set of observed flows. At the beginning we performed the non-parametric regression analysis that we have described in the previous chapter, and then we performed some test and scaling techniques. In the following section we will first apply this techniques to simulated data of a subset of 300 counties of the USA. As weights we use real population of each

locations (in $2000 - 2001$), and we use real data also for the number of people who leave a county. The deterrence function is stretched exponential

$$f(r) = e^{-r^{1.5}/1500}. \tag{5.5}$$

On these data we perform non-parametric regression and we use the results of the fit.

5.1 From P(r) to f(r)

When we perform the non-parametric fit, we need to start from reasonable values of the deterrence function $f(r)$ and weights, in order to possibly reach the global minimum of the cost function, and not only a local one. We have chosen to use as initial value of the deterrence function $P(r)$, i.e. the probability density function of length trip, because we expect it to have similar characteristics of that of $f(r)$.

In this section we describe a relation between $P(r)$ and $f(r)$ which can help to improve the efficiency of our algorithm for the non-parametric fit.

The fraction of trips with length r is:

$$P(r) = \sum_{i,j} T_{ij}\ P(T_{ij}|p_{ij}T_i)\ \delta(r_{ij} - r) / \sum_l T_l \tag{5.6}$$

where $P(T_{ij}|p_{ij}T_i)$ is the probability of observing T_{ij} trips from i to j. If we define $T = \sum_l T_l$, we have:

$$\begin{aligned} P(r) &= \frac{1}{T} \sum_{ij} T_i p_{ij} \delta(r_{ij} - r) \\ &= \frac{1}{T} \sum_i T_i \frac{\sum_j w_j f(r_{ij}) \delta(r_{ij} - r)}{\sum_k w_k f(r_{ik})} \\ &= f(r) \frac{1}{T} \sum_i \frac{T_i}{c_i} w_i(r) \end{aligned} \tag{5.7}$$

where $c_i \equiv \sum_k w_k f(r_{ik})$, and $w_i(r) = \sum_j w(r_{ij} = r)$ is the sum of all weights with distance r from i. At this point we can formulate two hypothesis:

1. c_i is independent from the origin i; if so, $c_i = c\ \forall i$.
2. $w_j(r)$ depend only on r; if so, $w_i(r) = w(r)\ \forall i$.

If these hypothesis are verified, we can write

$$P(r) = \frac{f(r)w(r)}{c}. \tag{5.8}$$

So we can reverse the formula to find $f(r)$ from $P(r)$, where f(r) is, as usual, defined apart from a constant:

$$f(r) = \frac{P(r)}{w(r)} \tag{5.9}$$

In practice, we calculate $w(r)$ as the average on i of $w_i(r)$. With this formula, it is possible to determine the deterrence function $f(r)$ from only $P(r)$ and the weights w_i, but we need to verify the hypothesis first.

For hypothesis 1, we compute $c_i \equiv \sum_k w_k f(r_{ik})$ for all i and plot the distribution $P(c_i)$, to see if it is peaked on an average value c. In figure 5.1 we can see the plot for synthetic data that we are considering.

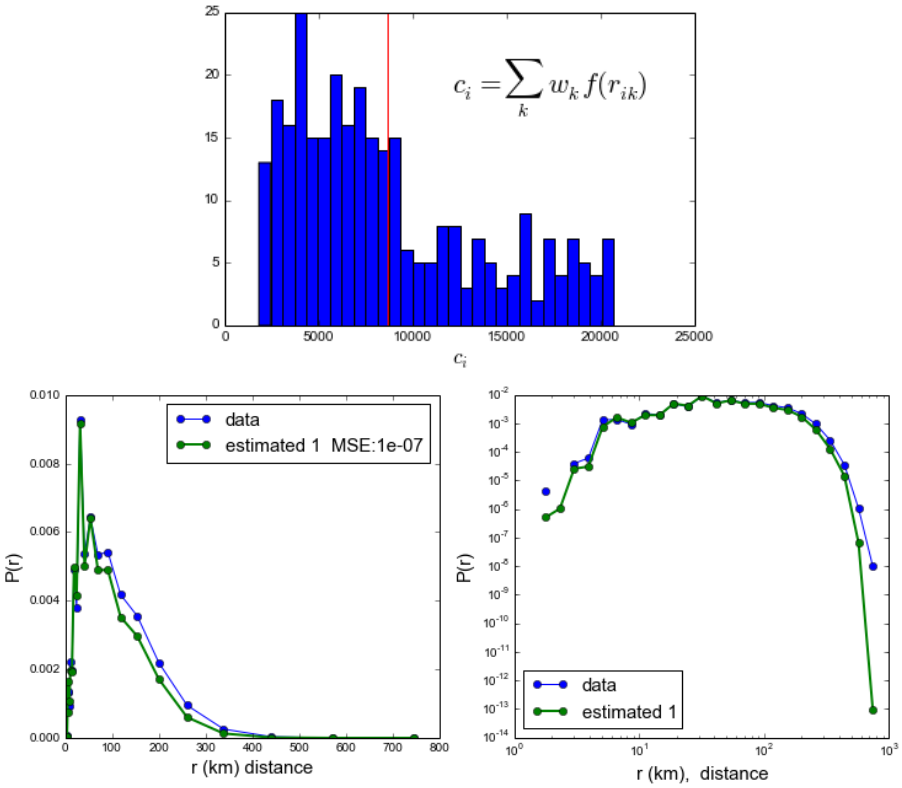

Figure 5.1: On the left, histogram of the values of c_i (with standard deviation: 0.594), which we approximated to be independent of i in the figure on the right of $P(r)$ using $P_c(r)$ (5.10).

Then we compute $P_c(r)$ using

$$P_c(r) = \frac{f(r)}{Tc} \sum_i T_i w_i(r) \tag{5.10}$$

to check if it is close to the real $P(r)$ (figure 5.1). Even though $P(c_i)$ is not particularly peaked (the ratio between mean and standard deviation is 0.594), $P_c(r)$ is a good approximation for $P(r)$. In general, the form of $P(c_i)$ depends on $f(r)$.

To verify the hypoteses 2, we compute $w_i(r)$ for every i at various r, plot the distribution $P(w_i(r))$, and check if each distribution, for a given r, is peaked. The distribution we find usually presents a peak and long tail on the right, so the ratio between the standard deviation and the mean is not particularly low. The hypoteses is verified well only for r big enough (relative errore is $\sim 1\%$ for $r \sim 90km$. $P(w_i(r))$ depends on how the weights are distribuited on the space, but on average we expect that $w(r)$ scales almost linearly with r, until the average radius of the area we are considering ($\langle r \rangle \sim 300km$), after which we expect a decrease. We can see this behaviour in figure (5.2). Then we compute $P_w(r)$ using

$$P_w(r) = \frac{f(r)w(r)}{T} \sum_i \frac{T_i}{c_i} \tag{5.11}$$

check if it is close to the real $P(r)$ (figure 5.2); we can see that the two function are similar apart from short distance ($r < 50$ km); this difference is due to anisotropy in distribution of cities for short distance. If we compute $P(r)$ with (5.8), the approximation is exactly the same

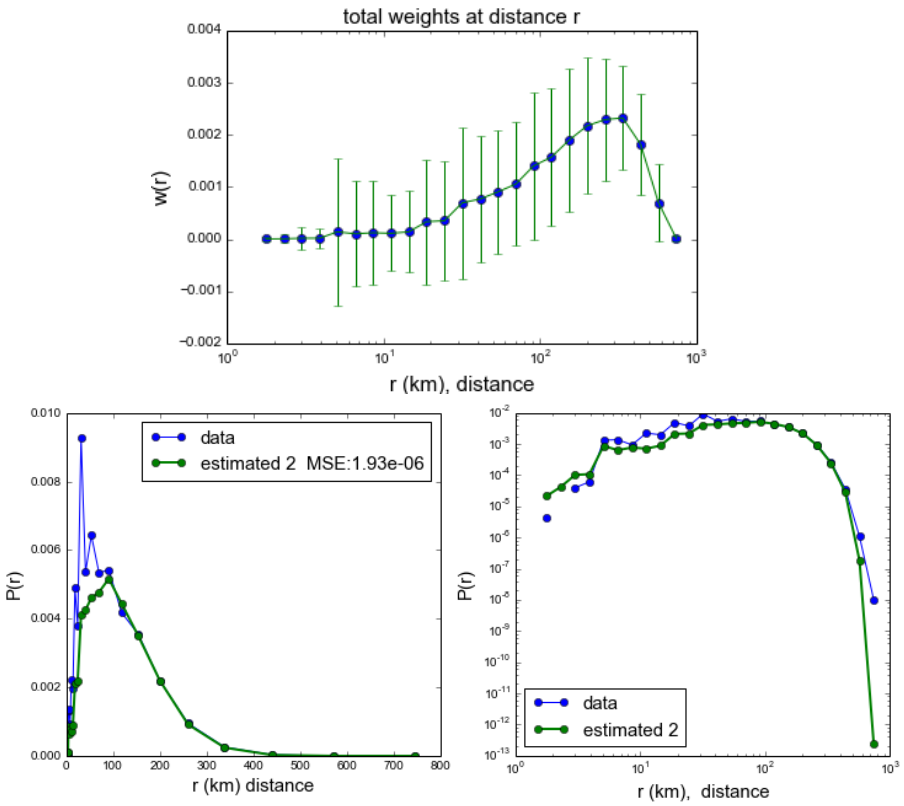

Figure 5.2: On the left, average values of total weights at distance r, with standard deviations. On the right, approximation of $P(r)$ using $P_w(r)$ (5.11), which assumes total weights at distance r from a location to be independent of the location.

as (5.11), apart from a factor that makes no difference after the normalization of the function. So actually the only important approximation that we need to make is $w_i(r) = w(r)$.

The value of $P(r)$ we obtain is quite in agreement with the function calculated from data, so we can use (5.9) to find an approximation for $f(r)$ and compare it with the deterrence function derived from the fit. As we can see in figure (5.3), the agreement is good for large r ($r > 60km$).

In this procedure we can also use population instead of weights, when we are dealing with real data (in this simulation we took the population as weights). There is an approximation due to binning, but the result does not change very much if we use smaller bin, furthermore we risk to have bin without any values. In our comparison, we have always used the same binning for deterrence function or pdf calculated with the different methods.

5.2 Probability Integral transform

We check that the multinomial assumption for $P(T)$ holds using the Probability integral transform (PIT). This technique allows to compare not only the average observed flows with the model's prediction, but also the entire distribution. It is a stronger test than plotting T_{data} vs T_{model}, and can be used also when the OD matrix is extremely sparse (few trips from each location) which is the case when one wants to study OD matrices at high spatial resolution.

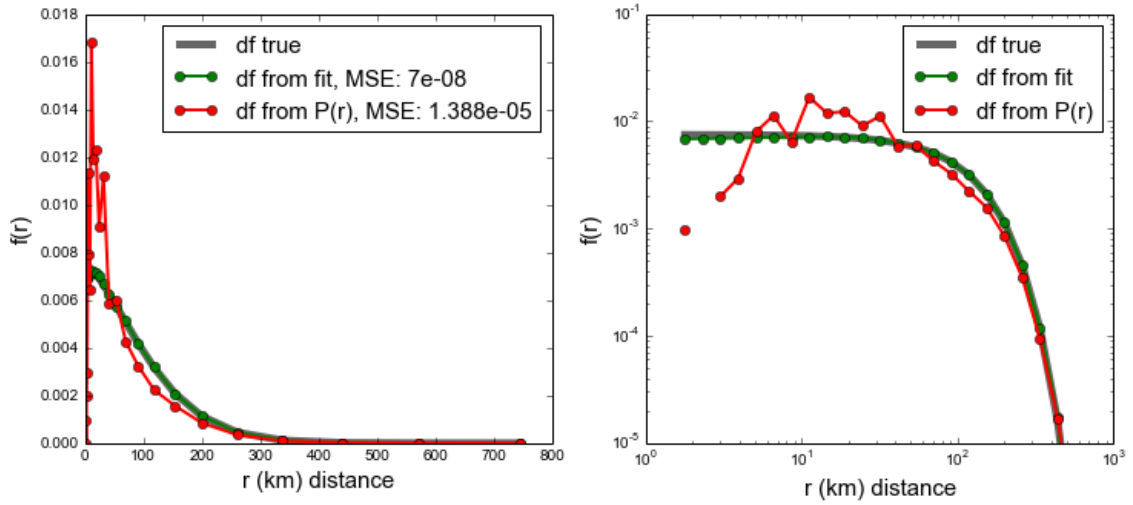

Figure 5.3: Approximation of $f(r)$ using (5.9) (which assumes total weights at distance r from a location to be independent of the location) compared with $f(r)$ derived from fit, and the true $f(r)$ that we used for generate data.

This method works on cumulative distributions and can be applied to data points generated by any number of different distributions.

The classic Probability Integral Transform Theorem can be stated as follows.

Theorem 1. *If a random variables X has a continuous distribution function F(x), then the random variables F(X) has a uniform distribution in the interval* $(0, 1)$*, that is, is a U(0,1) random variable.*

Let $X \sim F_X(x)$. Define the transformation

$$Y = F_X(X) \in [0, 1],$$

$$X = F_X^{-1}(Y),$$

here $\frac{dy}{dx} = F_X(x)' = f_X(x)$.

$$F_Y(y) = f_X[F_X^{-1}(y)] \frac{1}{f_X[F_X^{-1}(y)]} = 1,$$

i.e. Y is uniform over $[0, 1]$. Another way to see it is through the distribution function:

$$F_Y(y) = P(Y \leq y) = P(F_X(X) \leq y) = P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) = y.$$

The earliest use of this result was presented by Ronald Aylmer Fisher in his famous paper [Pro30] in 1930 where he introduced the theory of fiducial limit on intervals. Fisher again used the PIT in 1932 in the fourth edition of his book entitled "Statistical methods for research workers" [Fis32], in which he proposed a method to combine tests of significance.

The inverse Probability Integral Transform is used intensively in simulation of random variables.

In order to evaluate the goodness of fit we calculate the distance of probability integral transform obtained from the data, from the theoretical uniform one; we use different kind of measure, that we present at the end of this chapter. When we apply this technique to real data, we compare the distance that we find with the distance find with a simulated set of data with the same distribution of the real data.

5.3 Function collapse

In the study of critical phenomena, in Physics, three fundamental pillars are scaling, universality and renormalization [Sta99]. The scaling hypothesis was independently developed by several workers, including Kadanoff and Fisher. It has two categories of predictions, that are *scaling laws* and *data collapse*; for thermodynamic functions it is made in the form of a statement about one particular thermodynamic potential, generally chosen to be the Gibbs potential per spin, $G(H,T) = G(H,\epsilon)$ (H is the magnetic field, $\epsilon \equiv (T - T_c)/T_c$ is the reduce temperature, T_c is the critical temperature) . One form of the hypothesis is the statement that asymptotically close to the critical point, $G_s(H,\epsilon)$, the singular part of $G(H,\epsilon)$, is a generalized homogeneous function (GHF). Thus the scaling hypothesis may be expressed as a relatively compact statement that asymptotically close to the critical point, there exist two numbers, a_H and a_T (termed the field and temperature scaling powers) such that for all positive λ, $G_s(H,\epsilon)$ obeys the functional equation:

$$G_s(\lambda^{a_H} H, \lambda^{a_T}\epsilon) = \lambda G_s(H,\epsilon). \tag{5.12}$$

This equation constrains the form of a thermodynamic potential, near the critical point, so this constraint has implications for quantities derived from that potential, such as the equation of state. Starting from (5.12), it is possible to derive a way to scaling quantities, such as magnetization or temperature, in order that if we plot them, an entire family of different curves (like $M(H = const, T)$) will "collapse" onto a single function.

Inspired by this physical property, we look at the basic form of (5.1), in search for a relation between the variables that we expected to be present if (5.1) is a distribution in agreement with data. We find a function of flows and weights, that is different for every couple of locations. These different functions, after being scaled in a proper way, should collapse onto a single function common for all of them.

From (5.1) and $c_i \equiv \sum_k w_k f(r_{ik})$ we have:

$$f(r_{ij}) = \frac{p_{ij} c_i}{w_j}. \tag{5.13}$$

Using $p_{ij} = T_{ij}/T_i$, we find:

$$f(r_{ij}) = \frac{T_{ij} c_i}{T_i w_j}. \tag{5.14}$$

If we divide the data in bin and we plot $h(r_{ij}) = \frac{T_{ij}}{T_i w_j}$, we do not expect a good collapse of the function, because we have neglected c_i (see fig 5.4).

Let $g(r_{ij}, r_{ik})$ be a function define as follows:

$$g(r_{ij}, r_{ik}) \equiv \frac{f(r_{ij})}{f(r_{ik})}. \tag{5.15}$$

using (5.14):

$$g(r_{ij}, r_{ik}) = \frac{T_{ij}}{w_j}\frac{w_k}{T_{ik}}. \tag{5.16}$$

Then we procede as follow (we show here some results obtained from synthetic data):

- for all pairs of locations (i,j) and (i,k) we compute $g(r_{ij}, r_{ik})$ using (5.16). We exclude flows with less than 10 travellers because the statistics is too poor.

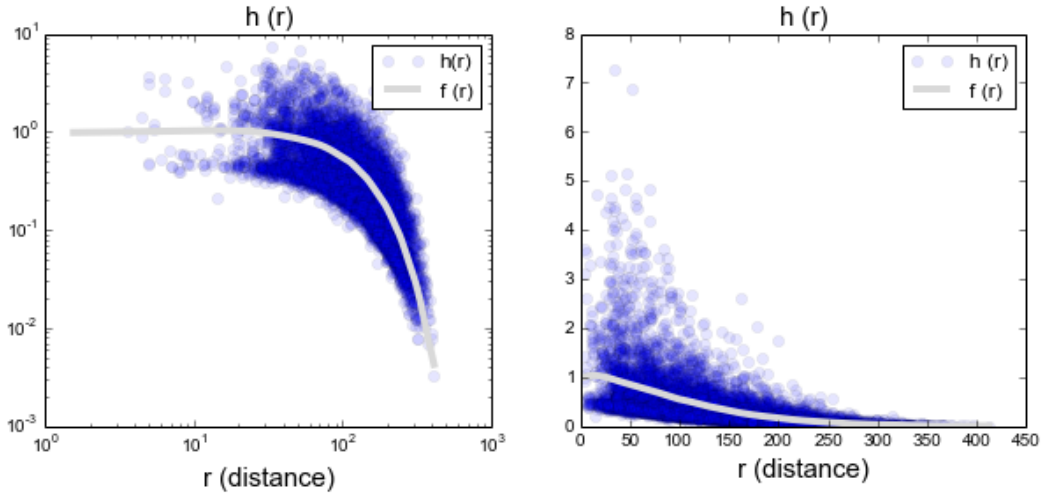

Figure 5.4: $h(r_{ij}) = \frac{T_{ij}}{T_i w_j}$ for hybrid simulated data. As we expected, with this function we do not see a good collapse.

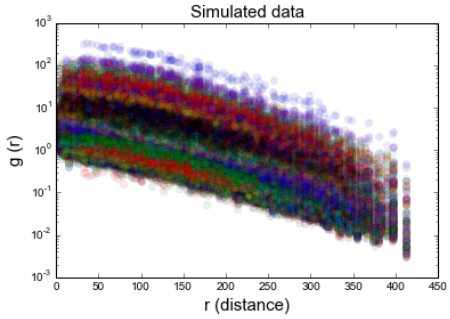

Figure 5.5: Hybrid simulated data, plot of the different curves $g_{ij}(r)$.

- After fixing r_{ij}, we divide r in bins and plot the curves $g(r_{ij}, r_{ik})$ for all bins. For semplicity, we can say that these curves represent the functions $g_{ij}(r)$, with $r = r_{ik}$. As expected, these curves are different from each other (fig 5.5).
- Consider two distances, r_{ij} and r_{lm}. Let $r_{ik} = r_{ln} = r$, we have the identity

$$g(r_{ij}, r_{ik}) = g(r_{lm}, r_{ln}) g(r_{ij}, r_{lm}). \tag{5.17}$$

  Plotting the two curves $g_{lm}(r)$ and $g_{ij}(r) = g_{lm}(r) g(r_{ij}, r_{lm})$, we should see a collapse (fig 5.5).
- Finally, after chosing a particular r_0, we plot all the curves $g_{ij}(r)$, $\forall i, j$, multiply by the right rescaling factor $g(r_0, r_{lm}) = f(r_{lm})/f(r_0)$ to see the collapse. In figure 5.6 we plot also $f(r)/f(r_0)$, that is the theoretical function where the other function should collapsed.

In order to evaluate the goodness of the collapse we calculate the distance of the collapsed curves from the theoretical one, using different kind of measure, that we present in the next section. When we apply this technique to real data, we compare the distance that we find with the distance find with a simulated set of data with the same distribution of the real data.

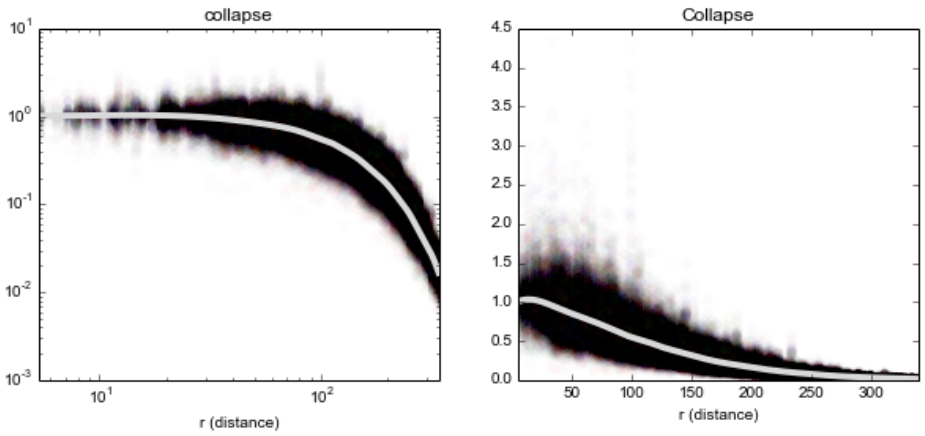

Figure 5.6: Collapse of the $g_{ij}(r)$ for hybrid simulated data. On the left: log-scale, on the right: linear scale. In white we represent the theoretical function $f(r)/f(r_0)$ where the other should collapse. Distance from the theoretical curve: jac 0.071, MSE 0.020, sor 0.102. We simulated data with the same distribution find from the fit, and we perform the same function collapse. The values of the distance that we find are very closed to the first ones: jac 0.068, MSE 0.018, sor 0.100.

5.4 Distance between probability density functions

To compare two probability density functions there are various distance/similarity measures that can be used [Cha07]. Here we list the distances that we chose for our analysis.

For two discrete probabilities distribution $P = (p_1, \ldots, p_d)$ and $Q = (q_1, \ldots, q_d)$, we can define:

- *Sørensen distance* (sor):
$$d_{sor} = \frac{\sum_{i=1}^{d} |P_i - Q_i|}{\sum_{i=1}^{d} (P_i + Q_i)}; \tag{5.18}$$
- *Jaccard coefficient* :
$$s_{jac} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}; \tag{5.19}$$
and *Jaccard distance* (jac):
$$d_{jac} = 1 - s_{jac} = \frac{\sum_{i=1}^{d} (P_i - Q_i)^2}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i}; \tag{5.20}$$
- *Kolmogorov-Smirnov* (ks):
$$d_{ks} = \sup_i |(P_i - Q_i)|; \tag{5.21}$$
- *Hellinger distance* (he):
$$d_{he} = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{d} (\sqrt{p_i} - \sqrt{q_i})^2}. \tag{5.22}$$

Chapter 6

Data analysis

Here we present the results of our analysis applied to data of commuting and migration flows in the USA and England.

The dataset which is better described by our model is the one which has distribution more similar to the theoretical one. We compare the distance between observed-estimate flows, the probability integral transform (the distance from the expected uniform distribution), the distance of the collapse from the expected distribution.

We will show that basic assumptions of singly-costrained Gravity Models are rather compatible with local and global commuting flows in England. The agreement is less good for commuting flows in the USA, and even worse for migration in the USA.We hypotesize that the wrong level of aggregation may be the reason of the discrepancies from model to data.

6.1 Datasets

Our Datasets consist of:

- US commuting
  Data on commuting trips between United States counties are available online at http://www.census.gov/population/www/cen2000/commuting/index.html. The files were compiled from Census 2000 responses to the long-form (sample) questions on where individuals worked. The files provide data at the county level for residents of the 50 states and the District of Columbia (DC). The data contain information on 34,116,820 commuters in 3,141 counties.
- US migrations
  United States population migration data from 1992-1993 to 2006-2007 are available online at
  http://www.irs.gov/uac/SOI-Tax-Stats-Migration-Data The main source of area-to-area migration data in the United States is the Statistics of Income Division (SOI) of the Internal Revenue Service (IRS), which maintains records of all individual income tax forms filed in each year. The Census Bureau is allowed access to tax returns, extracted from the IRS Individual Master File (IMF), which contains administrative data collected for every Form 1040, 1040A, and 1040EZ processed by the IRS. Census determines who in the file has, or has not, moved. To do this, first, coded returns for the current filing

year are matched to coded returns filed during the prior year. The mailing addresses on the two returns are then compared to one another. If the two are identical, the return is labeled a non-migrant. If any of the above information changed during the prior 2 years, the return is considered a mover. An aspect of this dataset is that flows with less than 10 people are not recorded for privacy reasons.

- England commuting
  Data of commuting flows in UK between 7201 wards are available at :
  http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcm

6.2 Local level: comparison of different dataset

The model that we want to test is the singly-costraint Gravity Model with weights and deterrence function as free parameters. Now we apply the methods presented in the previous chapters to compare the goodness of the fit of this model against our three datasets. We study the performance of the model on a regional level. The considered region is shown in figure (6.1).

For the same region we compare the data of migration of 2000-2001 (USAm) and the data of commuting flows in the USA (USAc). We make a comparison also with data of commuting flows in a region of 300 wards in England (UKc).

Our aim is to decide if observed flows are compatible with the assumption of a generic singly-costrained gravity model. To do so, we compare the various distributions that we find with the theoretical ones, and we calculate the distances: from observed to estimated flows (fig. (6.2), from probability integral transform to the expected uniform distribution (fig. (6.3)), from estimation for $P(r)$ and $f(r)$ and the function obtained from fit (fig. (6.5), fig. (6.6), fig. (6.7)). In the end we compare the distance of data collapse with the distance that we find with a simulation with the same parameter as the fit (fig. (6.4)); we perform a comparison also between the distance from the uniform distribution of probability integral transform, and the distance of simulated data (see caption of fig. (6.3)).

We can see that UK data are the best fitted by singly-costrained Gravity Model, then the fit of commuting flows is better than migration flows in the USA.

Which characteristics of the data are responsible for that? If we look at the number of non-zero flows in the data and we compare it with the number of non-zero flows predicted by our model (with the parameters of the fit) we find that data of migration in the USA as more zero flows that the other dataset (table (6.1)).

So it seems that the more non zero flows, the better is the fit. For (USAm) there is an order of magnitude of difference between non-zero flows and non-zero estimated flows. Actually data of migration does not report flows with less than 10 people (for privacy reason, because data derives from taxes). So we look at flows greater than ten observed (4750), and estimated (5009), and we conclude that the absence of flows with less than ten people in the data can explain the discrepancy between observed-estimated non-zero flows.

In (table (6.1)) we report also correlation weights-arrivals for real data, and the same correlations for synthetic data generated with the same distribution found from the fit of real

Table 6.1: Dataset, number of non-zero flows in the data, number of non-zero flows estimated from the fit, correlation weights arrival, average dimensions of locations (define as the average of the distance from each location to the closest one), average ranges, i.e. average travels distance. We can notice that in the USA travel range and locations dimensions are longer that in the UK.

data	non-zero	non-zero estimated	weights-arrival	average dim	average range
USA (USAm)	4750	29610	0.661 (0.688)	24.7 km	50.1 km
USA (USAc)	12322	31929	0.884 (0.894)	24.7 km	35.1 km
UK (UKc)	41239	48097	0.836 (0.800)	1.7 km	7.9 km

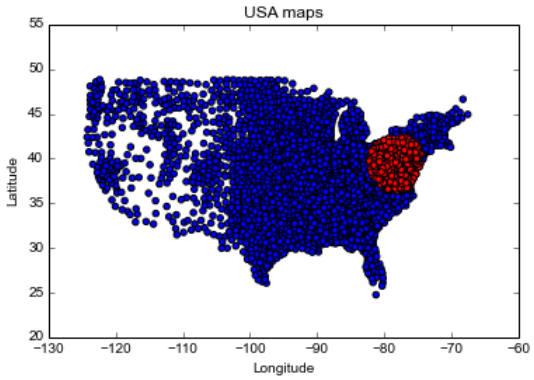

Figure 6.1: Region with 300 counties of the USA that we are studying.

data (in brackets in the table). We expect a strong relationship between weights-arrivals, like in simulated data of previous chapter, and this correlation is the reason why we chose arrivals as intial parameters of the weights for the non-parametric fit. Here we notice that the correlation is very similar from real to expected, but its value depends on the parameter. There is not a particular disagreement from real to simulated value; we notice though that commuting flows have a smaller correlation.

An aspect that distinguishes the datasets is that migration flows are characterized by a longer average travels distance (figure 6.2), and in general distances in the USA are longer than in UK. In the next section we will focus on the characteristics of deterrence function at local and global level.

The region of the USA that we study is the same that we studied in chapter 3 for (USAc) data. If we compare the distance of estimated flows with the real ones for the non-parametric fit of (USAc) with the results obtain with GLM models of chapter 3, we can see that this fit is definitely better. Taking into account the kind of data that we are studying, this model can be satisfactory, depending on the information that we are looking for. For example, it can tell us which is the best approximation for the deterrence function.

6.3 Deterrence function

When we look at the deterrence functions of figure 6.2 we notice that they have different characteristic.

The position of maximum depend on the scale, that is the average distance between

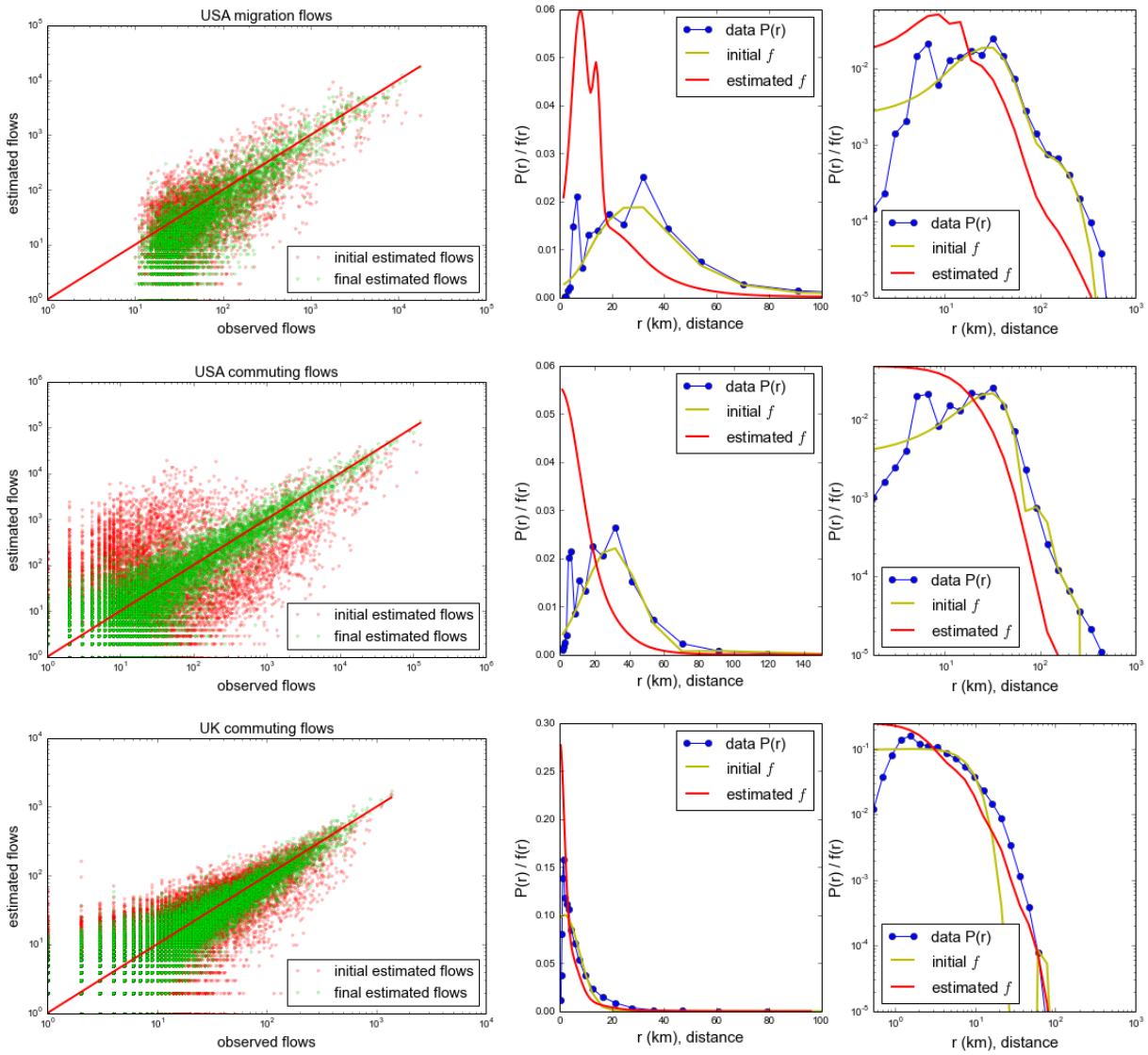

Figure 6.2: On the left: observed-estimated flows; on the right: P(r), initial deterrence function and optimal deterrence function. Data (from the top): USAm (migration USA), USAc (commute USA) and UKc (commute UK) in a region with 300 locations. We report chi-square and other distance between observed-estimated data:
USA migration: chi-square 471067, MSE 2759, jac 0.11 , sor 0.25.
USA commuting flows: chi-square 1168643, MSE 63605, jac 0.049, sor 0.16.
UK commuting flows: chi-square 86710, MSE 75.29966, jac 0.081 ,sor 0.19 .

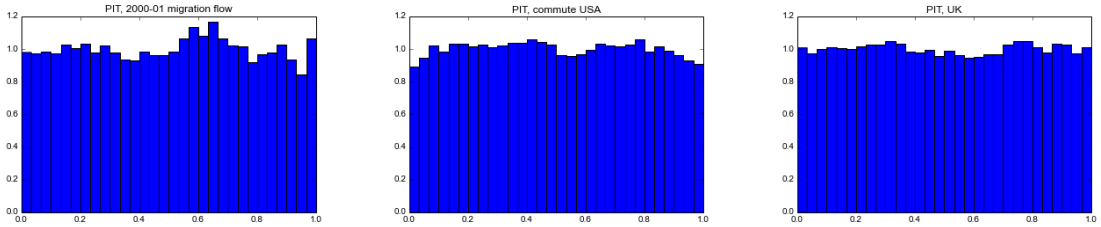

Figure 6.3: Probability integral transform and distance from the theoretical uniform distribution, for data (from the left): migration flows in the USA (USAm), commuting flows in the USA (USAc), commuting flows in England (UKc) .
Migration USA (USAm): MSE 0.00399 jac 0.0040 sor 0.024;
simulation: MSE 4.91e-05 jac 4.91e-05 sor 0.0026.
Commute USA (USAc):
MSE 0.00182 jac 0.0018 sor 0.018;
simulation: MSE 8.5e-06 jac 8.5e-06 sor 0.0012.
Commute UK (UKc)
MSE 0.00089 jac 0.00089 sor 0.013;
simulation: MSE 3.35e-05 jac 3.35e-05 sor 0.0022.

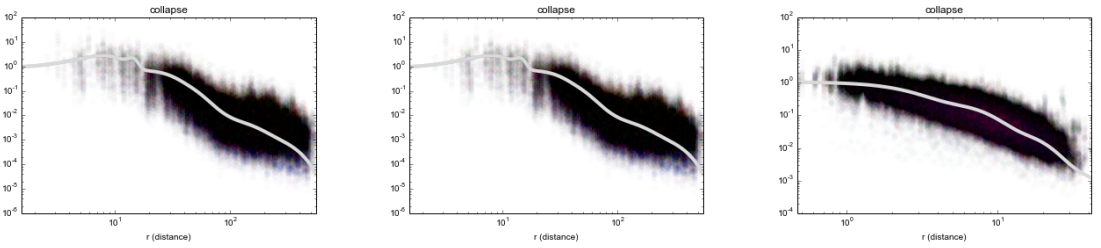

Figure 6.4: Function collapse for (from the left): migration in the USA (USAm), commuting flows in the USA (USAc), commuting flows in England (UKc). Here we report: the distance from the functions calculated from the data and the theoretical expected function and the distance from the functions calculated from the synthetic data (with the same distribution as the real data) and the theoretical expected function.
(USAm)
data: jac 0.632 MSE 0.136 sor 0.434 ks 12523 ed 48.1
simulation: jac 0.0398 MSE 0.00214 sor 0.081 ks 1869 ed 8.43
(USAc)
data: jac 0.8023 MSE 0.04916 sor 0.4749 ks 8088 ed 41.6
sumulation: jac 0.0321 MSE 0.000392 sor 0.0672 ks 911.5 ed 5.9
(UKc)
data: jac 0.33938 MSE 0.05343 sor 0.236 ks 45705 ed 67.1
simulation: jac 0.0637155 MSE 0.00540649 sor 0.103598 ks 17600 ed 27.8.

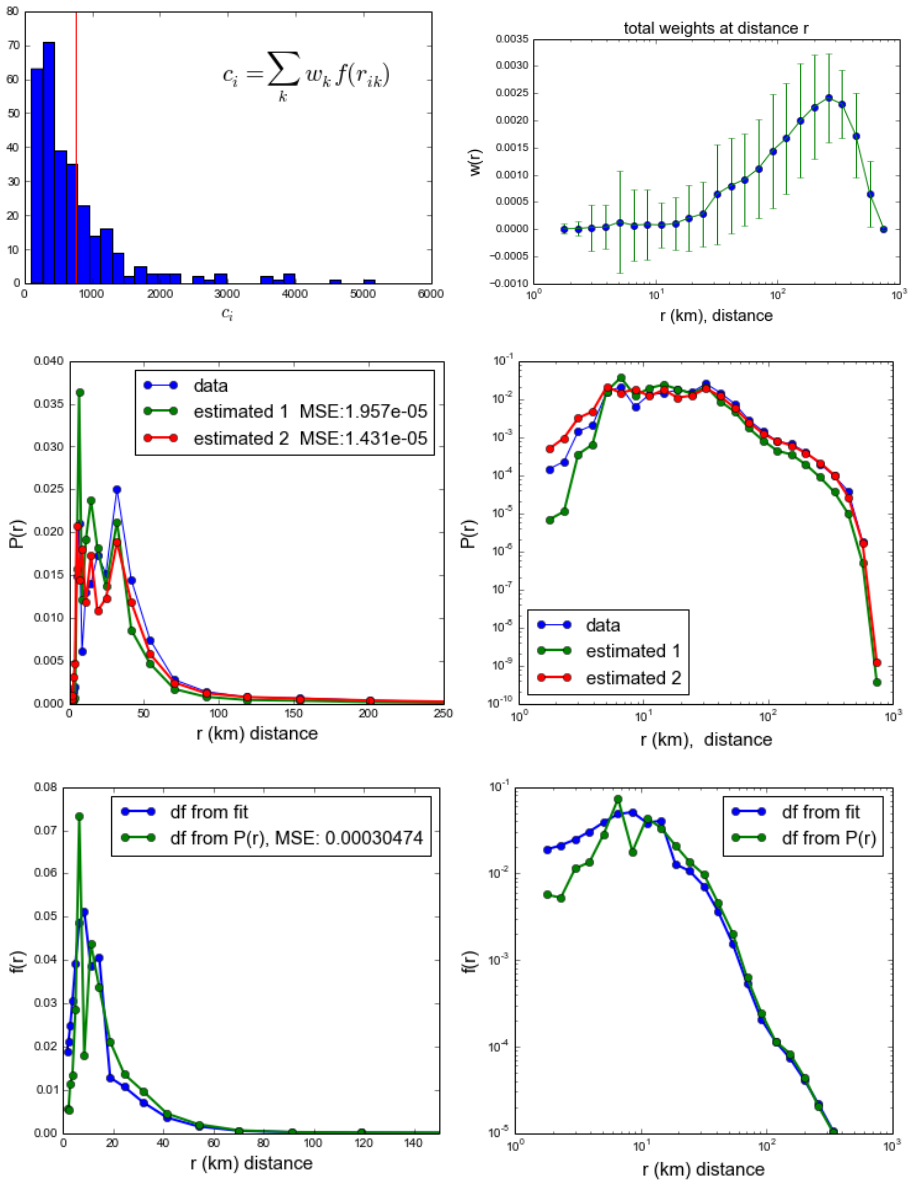

Figure 6.5: Data of migration flows in the USA (years 2000-2001) (USAm). On the top-right, the histogram of c_i, which we assume to be peaked in order to approximate $P(r)$ with estimation (1). On the top-left, average values of total weights at distance r from a location (with standard deviation), assumed to be independent of the location for estimation (2).Below , estimation (1) and (2) of $P(r)$; from the figure you can read the distance from real value to estimate of the $P(r)$, with the approximation $c_i = c$ (estimated 1), and with approximation $w_i(r) = w(r)\ \forall i$ (estimated 2). Below, estimated of $f(r)$ from $P(r)$ with approximation (1) and (2): in the figure you can read the distance from estimate of $f(r)$ (from the initial fit) and estimate of $f(r)$ from the $P(r)$ with the approximations.

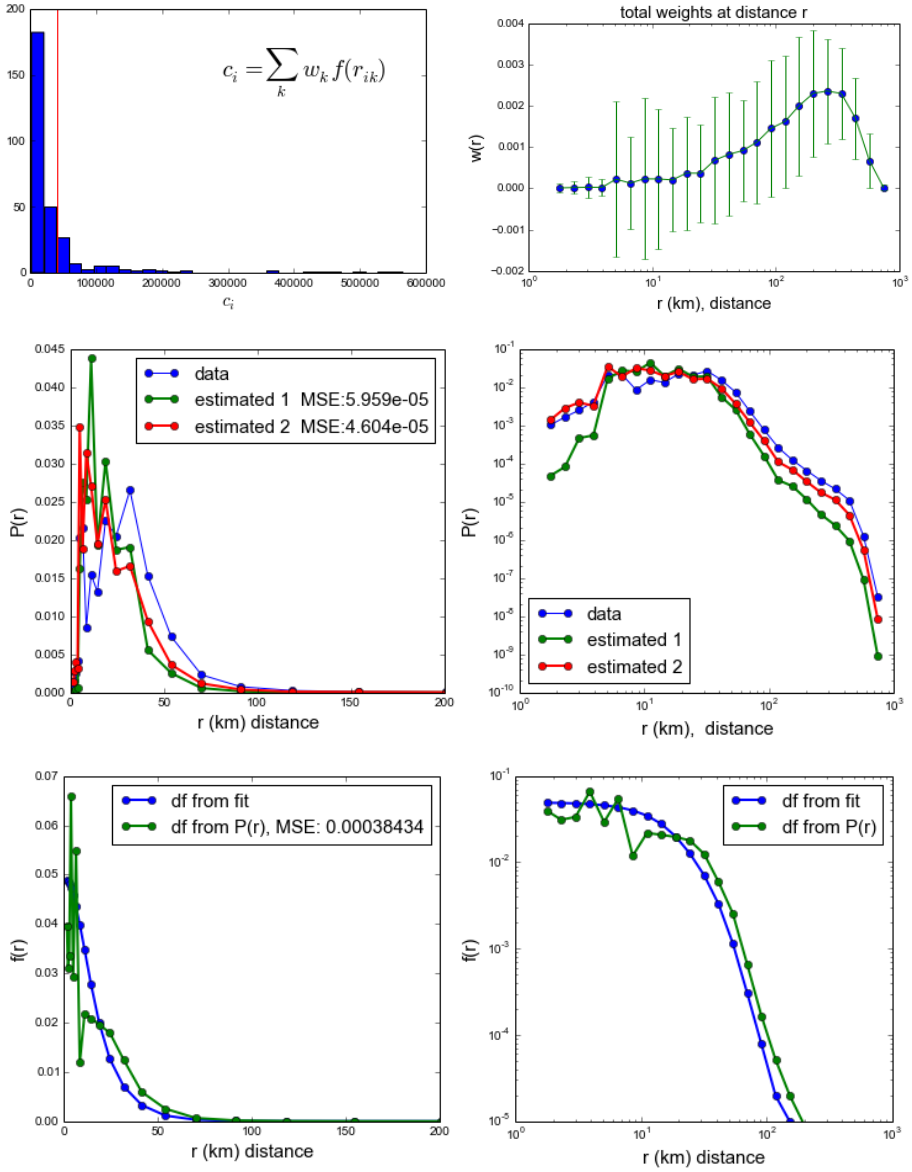

Figure 6.6: Data of commuting flows in the USA (years 2000-2001) (USAc). On the top-right, the histogram of c_i, which we assume to be peaked in order to approximate $P(r)$ with estimation (1). On the top-left, average values of total weights at distance r from a location (with standard deviation), assumed to be independent of the location for estimation (2).Below , estimation (1) and (2) of $P(r)$; from the figure you can read the distance from real value to estimate of the $P(r)$, with the approximation $c_i = c$ (estimated 1), and with approximation $w_i(r) = w(r)\ \forall i$ (estimated 2). Below, estimated of $f(r)$ from $P(r)$ with approximation (1) and (2): in the figure you can read the distance from estimate of $f(r)$ (from the initial fit) and estimate of $f(r)$ from the $P(r)$ with the approximations.

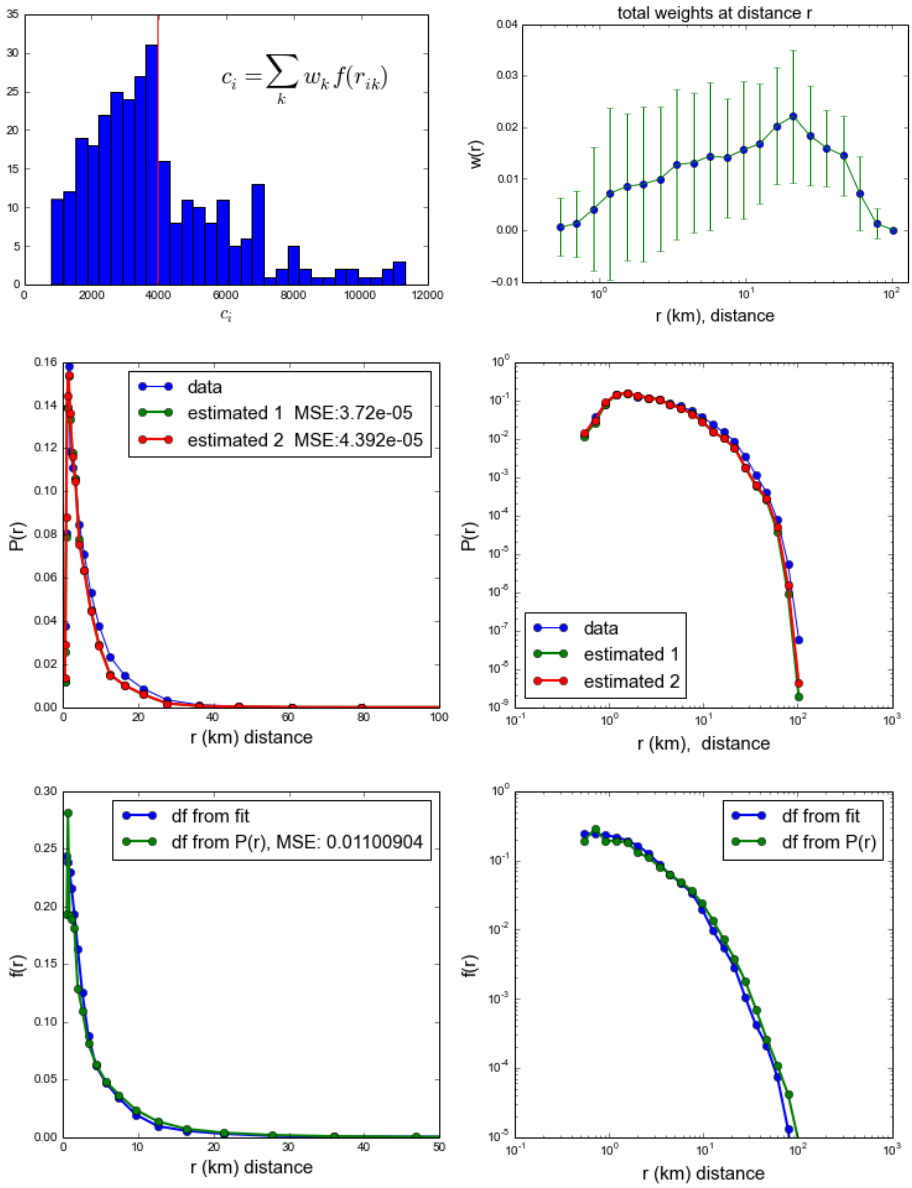

Figure 6.7: Data of commuting flows in the UK (UKc). On the top-right, the histogram of c_i, which we assume to be peaked in order to approximate $P(r)$ with estimation (1). On the top-left, average values of total weights at distance r from a location (with standard deviation), assumed to be independent of the location for estimation (2).Below , estimation (1) and (2) of $P(r)$; from the figure you can read the distance from real value to estimate of the $P(r)$, with the approximation $c_i = c$ (estimated 1), and with approximation $w_i(r) = w(r)\ \forall i$ (estimated 2). Below, estimated of $f(r)$ from $P(r)$ with approximation (1) and (2): in the figure you can read the distance from estimate of $f(r)$ (from the initial fit) and estimate of $f(r)$ from the $P(r)$ with the approximations.

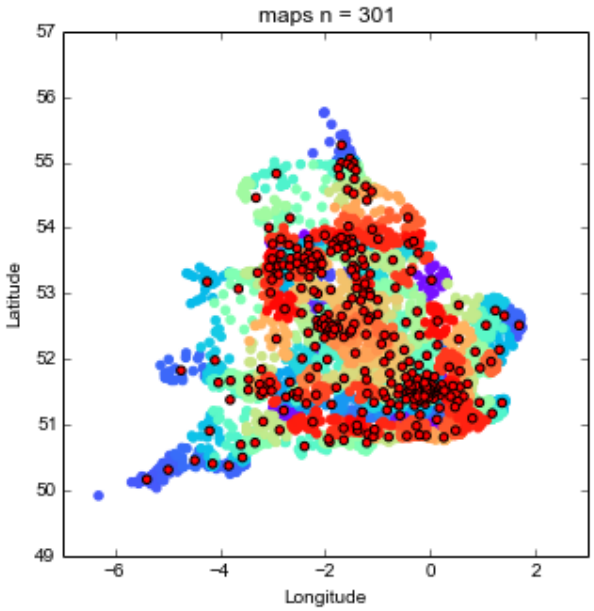

Figure 6.8: Aggregation in group of 24 wards in England and Wales. The different aggregation regions have different colors, and the red points correspond to their final position.

locations: the longer the distance, the further to the right the maximum of the function will be. In table (6.2) average linear dimensions derived from data are reported (locations dimension is the average of the minimum distance between locations).

Table 6.2: Average locations dimensions and average trip distance for data of after aggregation in group of 8 counties in USA, and 24 wards in the UK.

data	average locations dimension	average trip distance
USA (USAm)	110.8 km	670 km
USA (USAc)	110.8 km	173.1 km
UK (UKc)	13.6 km	27.6 km

Now we analyze the problem of scale. In the simulation of the previous chapter, if we compare the range of the deterrence funtion with aggregation and the range that we find in single region, both of them are around 300 km. This agreement is also due to the fact that the range is long and, even after aggregation, the minimum distance between counties (~ 80 km) is still shorter than the range.

It is interesting to see whether different regions share the same characteristics for the deterrence function, and if at local and global scales the range (and the peak) of deterrence function are the same. In figures 6.9 , 6.10 and 6.12, we present together the optimal deterrence functions found for different regions, and the function found from aggregation with fixed number of locations (8 for migration and commuting flows in the USA, 24 for England fig. 6.8). In the graphs of deterrence function, we always represent the function starting from the minimum distance values between pair of locations of the dataset.

When we fit separately different groups of 300 wards in England (fig. 6.9), we found that the characteristics of the deterrence function are in agreement with each other: range around

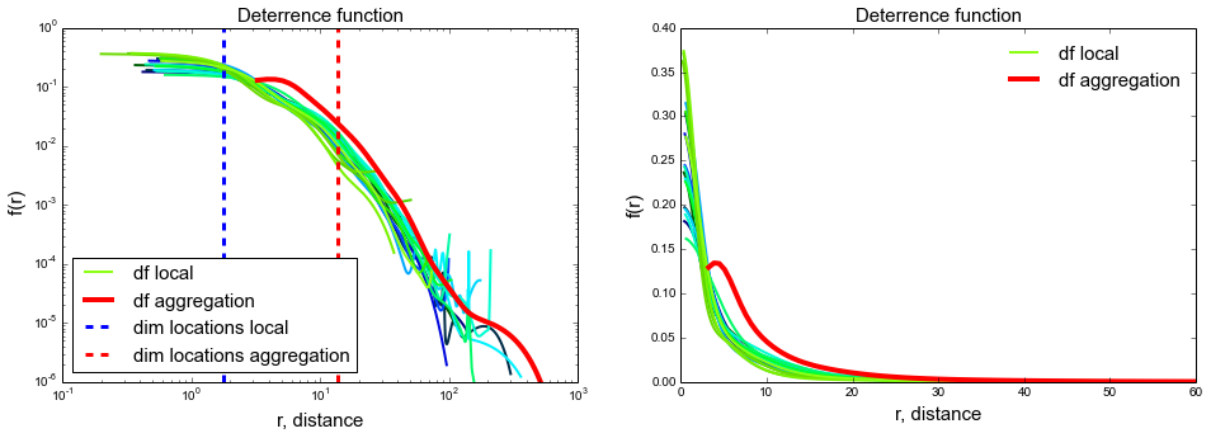

Figure 6.9: Deterrence function and $P(r)$ (logarithmic and linear scale) for England commuting flows. Results of the analysis in different regions, and after aggregation at global scale in fixed number of counties (24). Average dimension of locations (local level) : $\sim 1.8km$. Average dimension of locations (after aggregation) : $\sim 13.6km$.

40 km, just some noise for longer distance.

This behaviour tells us that England has homogeneous characteristic in the different areas. Furthermore, we see that the range is quite in agreement with that of aggregation. This is due to the fact that the shortest distances after aggregation are still shorter than the range of local deterrence function. In the next section we will perform the analysis to test the goodness of fit for the aggregated data of England.

When we aggregate counties in the USA, shortest distances after aggregation are already at the limit of the local range of the deterrence function, so we cannot expect the same regular behaviour that we find in England.

Despite that, in figure 6.10 we can see that there is less aggrement also between different region of the USA; the range is more or less the same, but the situation is less homogeneous and there is some noise for longer distance (100-1000 km).

Even though it can be partly due to the big distance after aggregation, still aggregation underline a definitely longer range for the deterrence function at global level ($\sim$ 140km, the double of the local values).

When we study data of migration, the situation is definitely eterogenous. In figure 6.11 we show the function at local level for three different areas (the three of them very central); they have various shapes, they all present a peak, and the range is around 80 km (but even for this there are fluctuations).

In figure 6.12 we show in more details the results of different aggregation numbers on migration in the USA. Despite the presence of noise, it is clear the position of the maximum and the range of the deterrence function: the peak is around 100 km, and then the range is around 400 km.

For migration in the the USA, the fit is clearly not good neither at local nor global level.

6.4 Test hypothesis on aggregated data

Now we apply scaling technique ad probability integral tranform to aggregation in fixed location number of synthetic data (group of 8 locations), USA commuting flows (group of 8

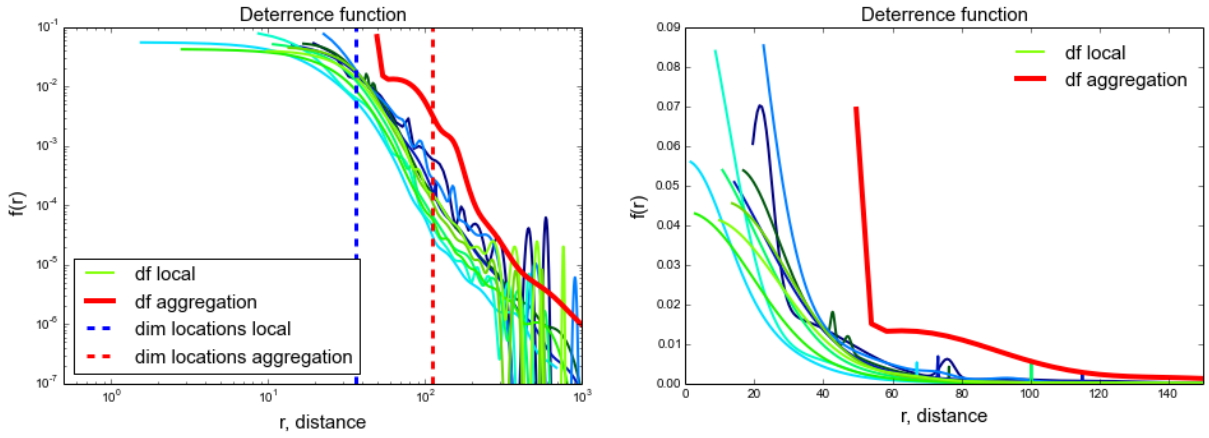

Figure 6.10: Deterrence function and $P(r)$ (logarithmic and linear scale) for USA commuting flows, 2000-2001. Results of the analysis in different regions, and after aggregation at global scale in fixed number of counties (8). Average dimension of locations (local level) : $\sim 36.5km$. Average dimension of locations (after aggregation) : $\sim 110.8km$.

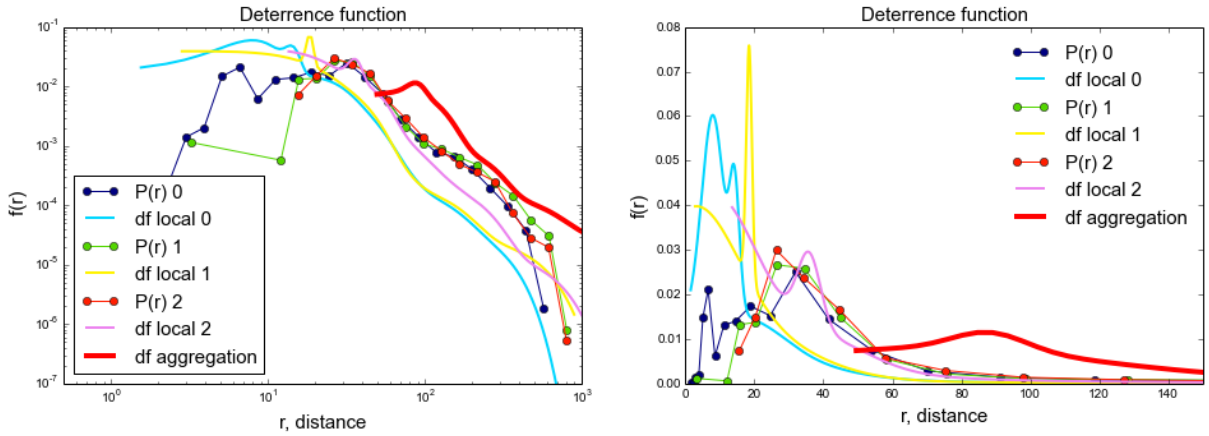

Figure 6.11: Deterrence function and $P(r)$ (logarithmic and linear scale) for migration in the USA (years 2000-2001). Results of the analysis in three different regions, and after aggregation at global scale in fixed number of counties (8).

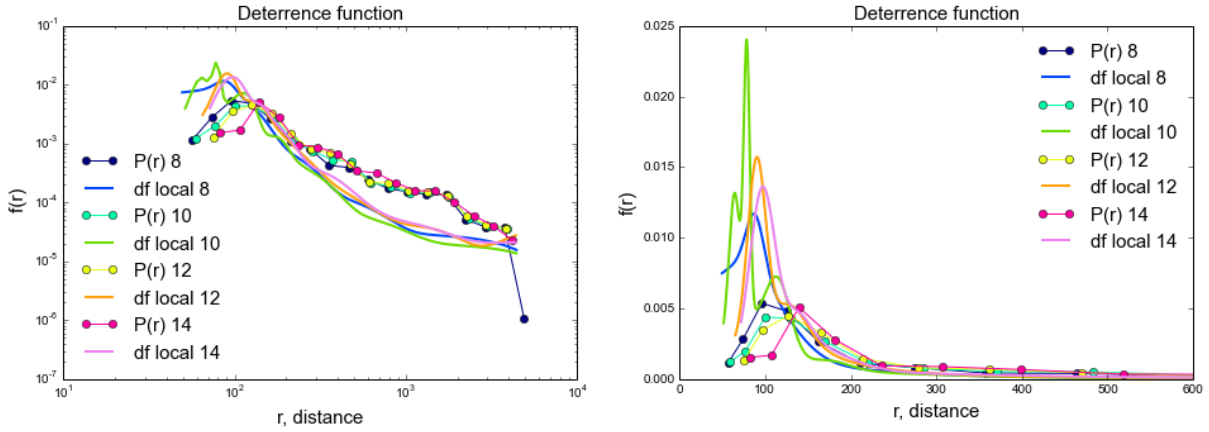

Figure 6.12: Deterrence function and $P(r)$ (logarithmic and linear scale) for migration in the USA (years 2000-2001). Aggregation in fixed group of counties (8, 10, 12, 14).

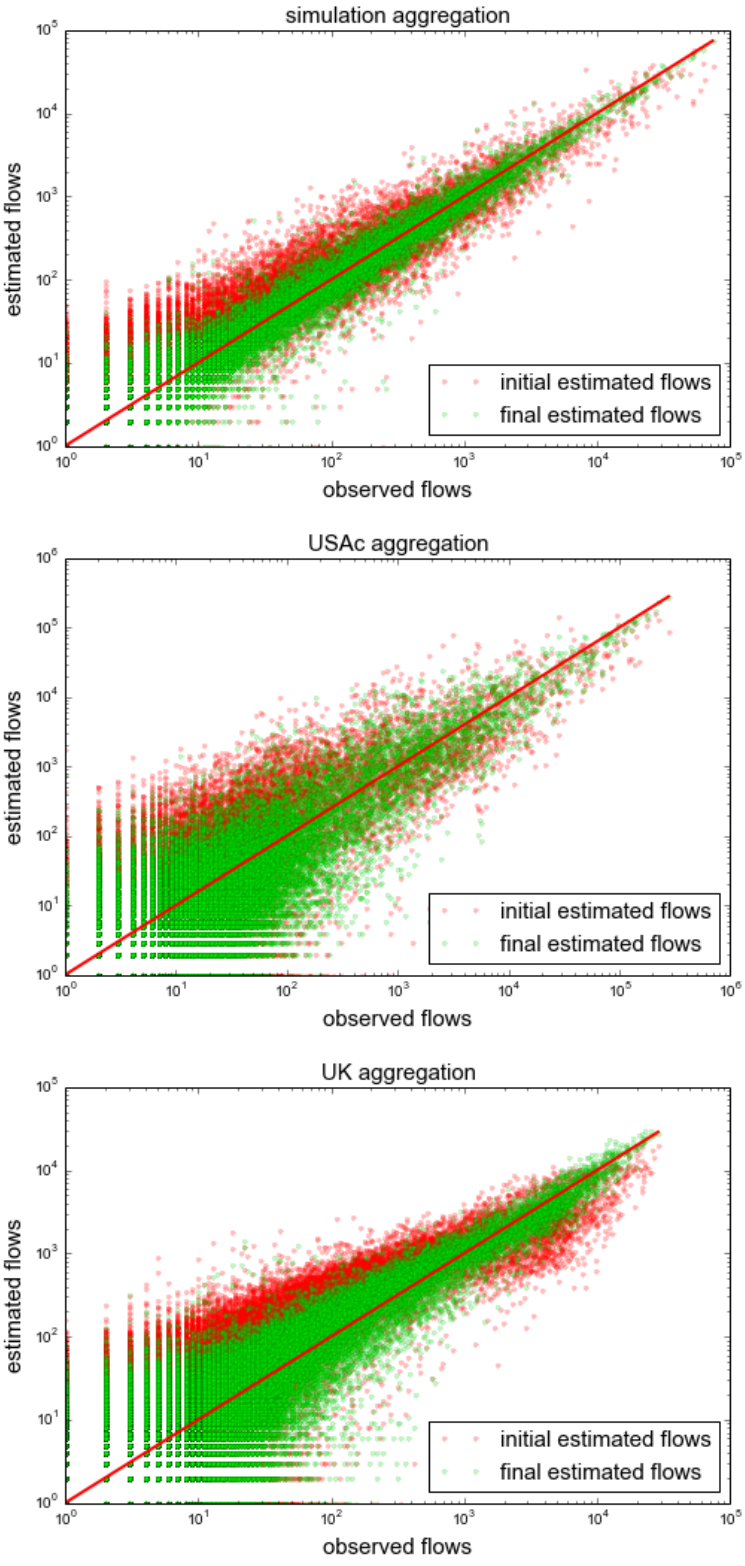

Figure 6.13: Observed-estimated flows for aggregate data. From the top: synthetic data (group 8), USA commuting flows, UK commuting flows.

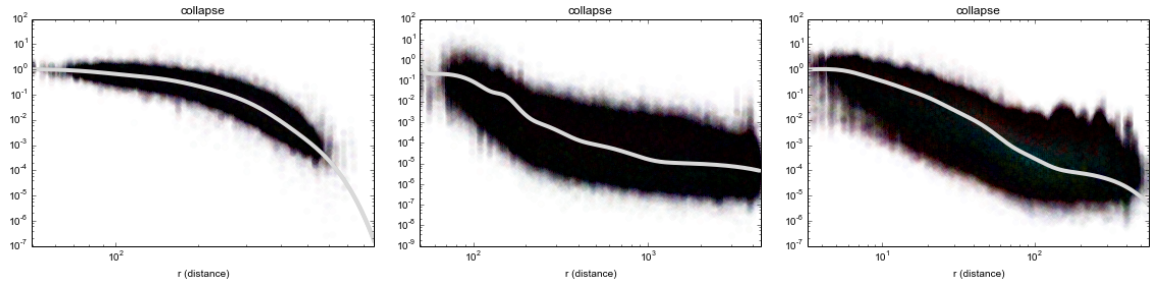

Figure 6.14: Function collapse for aggregate data. From the left: synthetic data, USA commuting flows, UK commuting flows.

locations) and UK commuting flows (group of 24 locations) (in fig 6.13 we show the plot of observed-estimated flows). When we speak about synthetic data, we refer to hybrid simulation (B) described in chapter 4, with real population as weights, real positions of locations, and stretched exponential as deterrence function (5.5).

In figure 6.14 we show the results of function collapse for aggregation in group of fixed location number, and in table (6.3) we report the distances. The aggregation is performed on synthetic data (group of 8 locations), USA commuting flows (group of 8 locations) and UK commuting flows (group of 24 locations). We compare the distance of the collapse with the distance of a simulated data with the same distribution. To understand the effect of aggregation, we start our analysis from aggregation on synthetic data to show that neither in the simulation the collapse is like in the direct simulation (without aggregation). We perform also probability integral tranform from the same aggregated systems, comparing the result with simulated data (fig 6.15, table(6.4)).

Table 6.3: Distance from data collapse to theoretical function. We report also the values for simulated data with the same distribution.

data	distances
Synthetic data	jac 0.230 MSE 0.0223 sor 0.1888 ks 27220 ed 47.4
simulation for synthetic data	jac 0.027 MSE 0.0018 sor 0.0618 ks 8889.6 ed 16.6
USAc	jac 0.940 MSE 0.0130 sor 0.6014 ks 10155 ed 50.6
simulation for USAc	jac 0.037 MSE 1.498e-05 sor 0.0727 ks 911 ed 5.72
UKc	jac 0.548 MSE 0.0144 sor 0.3805 ks 54710 ed 96.2
simulation for UKc	jac 0.022 MSE 0.000152 sor 0.056 ks 5968 ed 13.4.

From these results, we realize that the kind of discrepancy between the distribution that we find from data and the expected distibution, is comparable with the distance between aggregated synthetic data and their expected values!

From results in section 6.2 it seems that assumption of singly-costrained gravity model are not verified, in particular for commuting and migration flows in the USA. From these observations from aggregated flows, we hypotesize that assumption may be satisfied, and the discrepancy between real data from model may be due at the fact that we are looking at the spatial flows phenomena at an aggregation scale too long. In order to satisfying assumption, it may be necessary to look at migration and commuting flows at lower level of aggregation

Table 6.4: Distances from theoretical uniform distribution for Probability integral tranform. We report also distances calculated for simulated data with the same distribution of the results of the fit.

data	distances
Synthetic data (10) (311)	MSE 0.00071 jac 0.00071 sor 0.0109 chi-square 0.0212
simulation for synthetic data	MSE 2.34e-06 jac 2.341e-06 sor 0.000603 chi-square 7.02e-05
Synthetic data (8) (389)	MSE 0.00013 jac 0.00013 sor 0.0045 chi-square 0.0038
simulation for synthetic data	MSE 3.94e-06 jac 3.9405e-06 sor 0.00080176 chi-square 1.18e-04
USAc (389)	MSE 0.00271 jac 0.00270 sor 0.0235 chi-square 0.0813
simulation for USAc	MSE 1.32e-06 jac 1.323e-06 sor 0.000476 chi-square 3.97e-05
UKc (301)	MSE 0.00086 jac 0.00086 sor 0.0114 chi-square 0.0257
simulation for UKc	MSE 3.50e-06 jac 3.50e-06 sor 0.000784 chi-square 1.05e-04

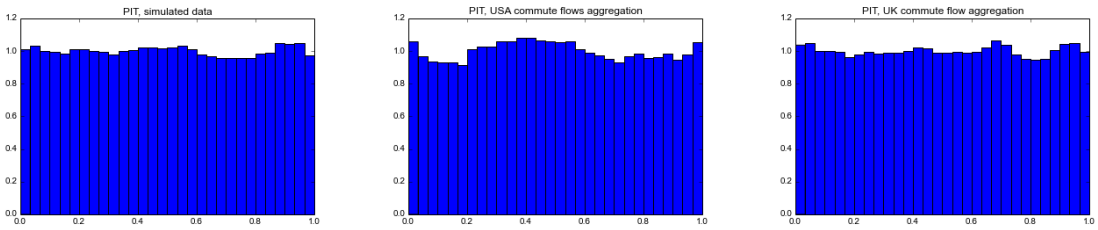

Figure 6.15: Probability integral transform for aggregate data. From the left: synthetic data (group 10), USA commuting flows, UK commuting flows.

respect than that of counties and wards. Wards in England are smaller than counties, and indeed commute flows in England are better described by the model. In addition to level of aggregation, also discretization should be chosen consistently with mobility processes.

6.5 Outliers

For data of commuting flows in a region of the USA, we look for outliers, i.e. locations that have flows on average farthest from the others respect to the predict value. As there is a relationship between population and estimated weights, we try to see if there are some locations whose optimal weights are farther from this relation. We find two locations with this behaviour, both with big population (fig. 6.16). We perform non-parametric fit excluding them and we find that distance between observed and estimated flows is better without the two locations, but just because we took off the biggest flows, as Probability integral transform is worst because there are less large flows. So we conclude that there are not locations which can be easily considered outliers.

In general the non-parametric fit can help us to find outliers. Another way to find outliers is to consider probability integral transform separately for each locations and consider the one with the longest distance from uniform distribution.

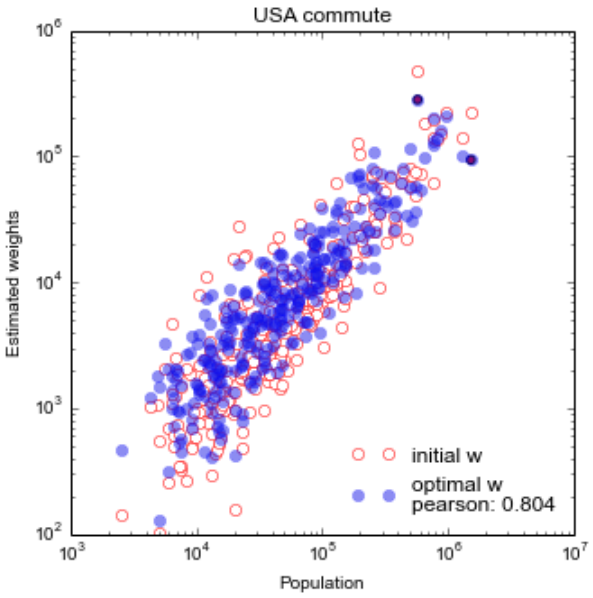

Figure 6.16: USA commuting flows; we underline the two counties which may be consider outliers.

Chapter 7

Conclusion

In this thesis we underline the common assumptions behind different formulations of the *singly-costrained* Gravity Model, and we develop a technique based on non-parametric regression to test these model assumptions.

The assumptions that we test are: independence of individual trips and probability of a trip proportional to a "weight" of the destination multiplied by a function of the distance (*deterrence function*).

We perform a non-parametric regression fit based on greedy algorithm. Non-parametric fit has the weights as free parameters, and uses a sum of ten gaussians to describe the deterrence function $f(r)$. The algorithm of the non-parametric fit consists of a minimization of a cost function (derived from likelihood) which describes the distance of observed values from values estimated by the model. The problem of a fit with many parameters (we are looking at sample of around ~ 300 locations plus 30 parameters of the sum of gaussians) is that we cannot be sure that we find the global minimum of the function. The idea behind the Greedy algorithm is to apply little change to a single parameter (randomly chosen) and then evaluate the new value of the cost function: only if it has decreased, the new parameter's value will be accepted. The initial values for the parameters are very important. To be sure to reach a reasonable solution, we start from arrivals as weights (because we find with simulation that there is a strong correlation between arrivals and weights). For the sum of gaussians, we start from a fit of $P(r)$, i.e., the probabilty distribution function of travel distance (derived from data). We choose it as initial value because it has reasonably similar characteristics to the deterrence function.

We notice that if some approximations hold, then there is a simple relation between $P(r)$ and $f(r)$. It is possible to derive $f(r)$ from the knowledge of $P(r)$ and weights. This can be a useful future improvement of the algorithm to perform a more efficient non-parametric fit: the minimization process would find only the optimal weights, and in every step of the algorithm the deterrence function would be derived from $P(r)$ and weights.

Starting from the results of non-parametric fit, we developed two techniques to assess the goodness of fit of the model: probability integral transform and data collapse. The latter of these draws inspiration from scaling in critical phenomena in Physics. We evaluate the distance between data and their theoretical distribution, and compare the distances with the ones obtained by simulation from the same parameters.

We apply this method to data of commuting and relocation flows in the USA, and commuting flows in England. Data of flows in the whole of the USA (or in the whole of England) are too numerous to apply a minimization algorithm ($\sim$ 3000 counties in the USA, $\sim$ 7000 wards in England). So we select a region with 300 locations in the USA (and in England as well), and we perform our method for testing model assumptions on three different kind of data (migration and commuting flows in the USA, commuting flows in the UK). Comparing the distances from the expected distributions, we conclude that commuting flows in England are better described by the model, but none of the three datasets is perfectly compatible with model assumptions.

To manage to study flows in the whole of the United States, we aggregate flows in order to have a number of locations accessible to our algorithm. We notice that the model performs worse after aggregation. We test model assumptions on the aggregate data of commuting flows in the USA and in England, and we compare the results with the one obtained from simulated data (with real position of counties and real population as weights). The discrepancies from distances obtained from data and the expected distances are comparable with discrepancies for aggregation on simulated data. So we conclude that the wrong level of aggregation may be the reason why the model assumptions are not perfectly followed by the datasets.

The advantages of non-parametric regression and scaling techniques are that it allows us to test the basic assumptions behind models, which is an important first step in the comprehension of phenomena. The non-parametric fit can also help in understading which aspects of the data model are less compatible. For example, whether it is possible to find outliers, which are locations with a behaviours which are particularly unexplained by the model.

A negative aspect of the non-parametric fit is that it is computationally difficult, and we manage to obtain good results only with a sample of around $300 - 400$ locations. So it can be used to study flows only between a limited number of locations.

To reach a better understanding of commuting and relocation flows, it would be interesting to improve the evaluation of the level of confidence of the validity of the assumptions. Furthermore, more work is necessary to understand the effect of aggregation and the importance of scale and discretization for the Gravity models of human mobility.

Bibliography

[Agr13] A. Agresti. *Categorical Data Analysis*. Wiley and Sons, University of Florida, Gainesville, Florida, 2013.

[Bar10] M. Barthélemy. Spatial networks. *Physics Reports*, 499(1):pp. 1–101, 2010.

[BHG06] D. Brockmann, L. Hufnagel, and T. Geisel. The scaling laws of human travel. *Nature*, 439:pp. 462–465, 2006. doi:10.1038/nature04292.

[Car65] H. C. Carey. *Principles of Social Science*, volume 3. J. B. Lippincott and Co., 1865.

[Cha07] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4):pp. 300–307, 2007.

[Dob08] A. J. Dobson. *An Introduction to Generalized Linear Models*. Chapman and Hall, 3 edition, 2008. 320 pp.

[Eas84] R. Eash. Development of a doubly constrained intervening opportunities model for trip distribution. *Chicago Area Transportation Study*, 1984.

[EEBL] P. Expert, T. S. Evans, V. D. Blondel, and R. Lambiotte. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences*.

[Erl80] S. Erlander. *Optimal Spatial Interaction and the Gravity Model*. Springer-Verlag, 1980.

[ES90] S. Erlander and N. F. Stewart. *The Gravity Model in Transportation Analysis: Theory and Extensions*. VSP, 1990. 226 pp.

[FA82] R. Flowerdew and M. Aitkin. A method of fitting the gravity model. *Journal of Regional Science*, 22(2):pp. 191–202, 1982. doi: 10.1111/j.1467-9787.tb00744.x.

[Fis32] R. A. Fisher. *Statistical Methods for Research Workers*. Edinburgh Oliver and Boyd, 4 edition, 1932.

[GHB08] M. C. Gonzáles, C. A. Hidalgo, and A. L. Barbarási. Understanding individual human mobility patterns. *Nature*, 453:pp. 779–782, 2008. doi:10.1038/nature06958.

[Gri] D. B. Grigg. E. g. ravenstein and the "laws of migration".

[HF84] K. E. Haynes and A. S. Fotheringham. *Gravity and Spatial Interaction Models*. SAGE Publications, 1984. 88 pp.

[JWS08] W. S. Jung, F. Wang, and H. E. Stanley. Gravity model in the korean highway. *Europhysics Letters*, 81(4):48005, 2008.

[KCRB09] G. Krings, F. Calabrese, C. Ratti, and V. D. Blondel. Urban gravity: a model for intercity telecommunication flows. *Journal of Statistical Mechanics*, 2009(7), 2009.

[KKGB10] P. Kaluza, A. Kölzsch, M. T. Gastner, and B. Blasius. The complex network of global cargo ship movements. *J. R. Soc. Interf.*, 7:1093–1103, 2010.

[NW72] J. A. Nelder and R. W. M. Weddemburg. Generalized linear models. *Journal of the Royal Statistical Society*, 135(3):pp. 370–384, 1972. doi:10.2307/2344614.

[Pro30] Proceedings of the Cambridge Philosophical Society. *Inverse Probability*, volume 26, 1930. pp. 528-535. R. A. Fisher.

[Rav85] E. G. Ravenstein. The laws of migration. *Journal of the Statistical Society of London*, 48(2):pp. 167–235, 1885. doi: 10.2307/2979181.

[Rui67] E. R. Ruiter. Toward a better understanding of the intervening opportunities model. *Transportation Research*, 1(1):pp. 47–56, 1967. doi:10.1016/0041-1647(67)90094-9.

[Sch59] M. Schneider. Gravity models and trip distribution theory. *Papers in Regional Science*, 5(1):pp. 51–56, 1959. doi: 10.1111/j.1435-5597.1959.tb01665.x.

[SGMB12] F. Simini, M. C. González, A. Maritan, and A. L . Barabási. A universal model for mobility and migration patterns. *Nature*, 484(96):pp. 191–202, 2012. doi:10.1038/nature10856.

[SKWB10] C. Song, T. Koren, P. Wang, and A. L. Barabási. Modelling the scaling properties of human mobility. *Nature Physics*, 6:pp. 818–823, 2010. doi:10.1038/nphys1760.

[SS95] A. K. Sen and T. E. Smith. *Gravity Models of Spatial Interaction Behavior*. Springer London, 1995. Limited.

[Sta99] H. Eugene Stanley. Scaling, universality and renormalization: Three pillars of modern critical phenomena. *Reviews of Modern Physics*, 71(2):pp. 358–366, 1999.

[Sto40] S. A. Stouffer. Intervening opportunities: A theory relating to mobility and distance. *American Mathematical Society*, 5(6):pp. 845–867, 1940. doi:10.2307/2084520.

[TTG+10] C. Thiemann, F. Theis, D. Grady, R. Brune, and D. Brockmann. The structure of borders in a small world. *PLoS ONE 5, e15422*, 5(11):e15422, 2010. DOI: 10.1371/journal.pone.0015422.

[Wil69] A. G. Wilson. The use of entropy maximising models in the theory of trip distributions, mode split and route split. *Journal of Transport Economics and Policy*, pages pp. 108–126, 1969.

[Zip46] G. K. Zipf. The p1 p2/d hypothesis: On the intercity movement of persons. *American Sociological Review*, 11(6):pp. 677–686, 1946. 10.2307/2087063.