Imperial College London

Spatial Scaling in Human Mobility Models: a framework for analysing mobility flows across different spatial scales

CID: 00938658

Department of Physics, Imperial College London

30^{th} April 2018

Project Code: THEO-Evans-2

Supervisor: Dr. Timothy Evans
Assessor: Prof. Ray Rivers

(9,989 words)

Spatial Scaling in Human Mobility Models

Abstract

Human mobility models are mathematical frameworks that capture and predict the statistical properties of the movements of people, at the individual or collective level. Despite their significant success in the predictive analysis of commuting and migration flows, the spread of diseases and other aspects of human dynamics, all models are tailored to either small (cities) or large (countries) length scales. Therefore no unique framework has yet been developed to provide an accurate description of mobility flows at multiple spatial scales.

Here, we conduct a study of the scale-dependence of the gravity and the radiation law, and find a measure of the error introduced in the estimated flow when rescaling a system through a renormalisation procedure.

We illustrate our framework through simulations using a toy model and the real population distribution of London and Birmingham and observe strong agreement between our analytical predictions and the simulated results. We find that the gravity model is generally more robust to changes in the spatial scale, provided the deterrence function is correctly calibrated to a given coarse-graining level.

We suggest a quantitative approach to minimise the scaling error and provide a functional relationship of the distance exponent in the power-law gravity model.

Acknowledgments

First, I would like to express my gratitude to my project supervisor, Dr.Timothy Evans, for providing such an interesting and open-ended project. His guidance and support have been crucial in carrying out this research and the weekly meetings have always been full of stimulating discussions and pointed us in interesting and new directions.

I would also like to thank my project partner for the great spirit of collaboration and all the fruitful conversations we have had throughout the year.

Finally, I would like to thank Vaiva Vasiliauskaite, who was kind enough to share her experience as a doctoral student to help me draft this report.

Contents

Chapter 1

Introduction & Overview

1. Human Mobility

The study of human mobility is of primary importance in understanding many social and biological processes: as extensively reported in the literature, statistics on human movements are pivotal in explaining the distribution of economic activities, the spread of information, congestion patterns, the structure of cities and the propagation of infectious diseases [1].

Starting from the second half of the 19^{th} century, attempts to frame mobility phenomena such as migration [2] within a formal setting gave rise to the first quantitative treatments of human mobility. Following these early efforts, the last century witnessed the development of several mathematical models that attempt to refine our understanding of movements at both the individual and the population level, together with the mechanisms that drive them. These models have proven accurate to various degrees and continue to serve both practical and scientific purposes, as they can have far-reaching implications in a broad range of contexts, such as urban planning [3, 4], the design of public transport infrastructures, epidemiology [5, 6], emergency response [7] or archaeology [8], among others. These frameworks are particularly useful in data-scarce contexts, where their ability to predict human flows can be used as a forecasting or planning tool. Here we are exclusively concerned with population-level models, whose main aim is to capture the aggregate movement of people by estimating the statistical distribution of trips between spatial units.

1.1 The role of spatial scale in human mobility

The increasing availability of digital traces provided by GPS and mobile phone records allows us to easily obtain geographical information and gain new insights from the study of spatial interactions between people. Owing to the heterogeneity of the data granularity and the multiscale nature of human interactions, which can span several orders of magnitude, any reasonable framework that aims to describe and predict mobility patterns should be both universal and scale-invariant [1].

Far from being merely a desirable feature, the property of universality may also be necessary on theoretical grounds to provide a coherent picture of phenomena of human interactions at all spatial scales. This conjecture is further corroborated by the several examples of scaling behaviour of human phenomena at the individual level [9, 10]. There exists in fact a vast body of literature within the fields of geography, social sciences and complexity science, dedicated to uncovering the scaling laws characterising spatial interactions of people, ranging from the traditional Zipf's law [11] to the more recent discoveries of scaling relations characterising movement in cyberspace [12].

At the population-level, however, the scaling properties of human mobility are, perhaps surprisingly, still largely unexplored. The majority of the macroscopic models proposed so far are in fact tailored to the description of mobility patterns within specific spatial ranges [13]. This results in a different treatment of the flows depending on the length scale of interest and the models therefore fail to accurately reproduce mobility patterns across a wide range of scales. Moreover, as noted by Barthélemy, understanding the interplay of different spatial scales in human interactions represents an interesting open question, from a purely theoretical perspective [14]. Remarkably, the recently increasing interest in the study of scaling properties within human laws has prompted the development of new human mobility models that are intrinsically scale-invariant [15]. However, a strong validation of these models on large sets of data is still missing and traditional models are far from being replaced as the main modelling mechanisms.

Finally, despite a number of attempts to address the issue of the suitability of traditional models at varying spatial scales [13, 16], so far these comparisons have mostly been carried out with reference to commuting data or other empirical trip distributions. In contrast, we seek a more complete and data-agnostic approach, in which the scale dependence of the models is probed by analysing their mathematical form and performing simulations to validate our analysis. Ultimately, this project is motivated by the lack of an adequate investigation of the role of spatial scales on models of spatial interactions at the population level.

2. Human Mobility Models

2.1 General framework for mobility flows

We define a *spatial system* as a collection of units embedded in two-dimensional Euclidean space. These spatial units can be administrative units (Chapter 4, Sec. 2) – boroughs, cities and counties – artificial units – sites in a synthetic population (Chapter 3) – or clusters of any of these types of units (Chapter 4). We model the interactions within a spatial system using the standard framework of flow models. These refer to the movement of entities – people, diseases, banknotes, ideas, etc. – as a *mobility flow*. A common way to characterise space in order to analyse mobility flows within a system is to partition the area of interest into cells and then define an origin-destination matrix (ODM), whose T_{ij} entries correspond to the number of people moving from cell i to cell j per unit time. This allows us to treat aggregated mobility at different resolutions: starting from the maximal resolution provided by the data (real or synthetic), the partitioning can be arbitrarily chosen to fit the spatial scale of interest. The resulting OD matrix therefore encapsulates all the information about the flows at the considered resolution. Although this is not always necessarily the case, we consider here the same spatial units to be both origin and destination locations, thus obtaining an $N \times N$ OD matrix. By convention, self-loops are not considered, so $T_{ii} = 0, \forall i$ and the OD matrix is a hollow square matrix characterising the aggregate flows between units

$$\mathbf{T} = \begin{bmatrix} 0 & T_{12} & \dots & T_{1n} \\ T_{21} & 0 & \dots & T_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ T_{n1} & T_{n2} & \dots & 0 \end{bmatrix}. \tag{1.1}$$

Within the aforementioned flow models, two different schools of thought have dominated the research landscape, one which argues that mobility is directly hindered by geographical distance and one that instead ties the probability of a trip to the presence of higher *benefits* in the surrounding area. The former is represented by the long-standing gravity model, popularised in its modern form by Zipf [11], while the latter was introduced through the recent radiation model, proposed by Simini *et al.* in 2012 [17]. These two models have extensively been compared in the literature [13, 18, 19]. In fact, although by no means exhaustive, the analysis and comparison of these two frameworks provides a sufficiently complete overview of the most common advantages and limitations of the spatial modelling of population flows. However, so far most of these studies have been based on the analysis of their predictive performance against mobility patterns observed in empirical data but do not provide sufficient insight into the fundamental difference between them. For this reason, and given their popularity and diverging approaches, we exclusively focus on these two modelling paradigms and conduct an analysis through mathematical reasoning and numerical simulations with the aim of bridging the gap between the observed dissimilarities and our understanding of their mathematical structure.

In the two sections below, their mathematical formulation is reviewed, with an emphasis on their treatment of spatial scales and the main conventions used in this report are established.

2.2 The gravity model

The gravity model, in its most general form, hinges on the assumption that the mobility flow T_{ij} from site i to site j is governed by a law similar to that of the gravitational interaction between two bodies. That is, this interaction is proportional to the product of their respective populations m_i and m_j and decays with the distance r_{ij} between them. The number of trips T_{ij} between i and j is then predicted to be

$$T_{ij} \propto m_i m_j f(r_{ij}). \tag{1.2}$$

In this framework, the population is used as a proxy for the attractiveness of the location and the cost of travelling is encapsulated by the *deterrence function* f, which expresses the relationship with distance and most commonly takes the form of an exponential or power law decay [20]:

$$f_e(r_{ij}) = e^{-\gamma r_{ij}} \tag{1.3a}$$

$$f_p(r_{ij}) = r_{ij}^{-\gamma}, \tag{1.3b}$$

where γ is the distance exponent. Although various definitions of distance (time of travel, social distance, road distance) can and have been used [18], in the context of this report, we always refer to the Euclidean distance between two spatial units.

In order to create a general framework and compare the predicted flow across different models on an equal footing, our aim is to obtain a trip distribution law or, in other words, express the flow as the *relative* number of trips originating in i and terminating in j. To accomplish this, we rewrite Eq. (1.2) as

$$T_{ij} = O_i p_{ij}, \tag{1.4}$$

where

$$O_i = \sum_j T_{ij} \tag{1.5}$$

is a constant for each unit i and p_{ij}, representing the probability that an individual located at i travels to j, is subject to the constraints

$$\sum_j p_{ij} = 1, \qquad p_{ij} \geq 0. \tag{1.6}$$

This is referred to as the *production-constrained* version of the gravity model [21], in which the total number of trips "produced" by each site, the *outflow* O_i, is fixed. For simplicity, throughout this paper we take

$$O_i = m_i, \tag{1.7}$$

so that all individuals within the population of i have a non-zero equal probability of travelling. Although of course real systems depart from this simplification because of differences in social structures, it is reasonable to assume that the outflow of a given location is at least proportional to its population. Moreover, since, for the purpose of our analysis, only relative flows are relevant (Chapter 2), turning this proportionality into an identity can be done without loss of generality.

This hence allows us to obtain the trip distribution law for the gravity model:

$$p_{ij} = k_i m_j f(r_{ij}), \tag{1.8}$$

where, to satisfy the marginal constraint in Eq. (1.5), we have introduced the normalisation factor k_i, which takes the form

$$k_i^{-1} = \sum_j m_j f(r_{ij}). \tag{1.9}$$

It is worth noting that the presence of the distance parameter γ requires that the deterrence function be calibrated to the observed dataset. This is typically done through a regression analysis [22] and can yield different results even within the same dataset. This inconsistency has important implications for the physical interpretation of the deterrence

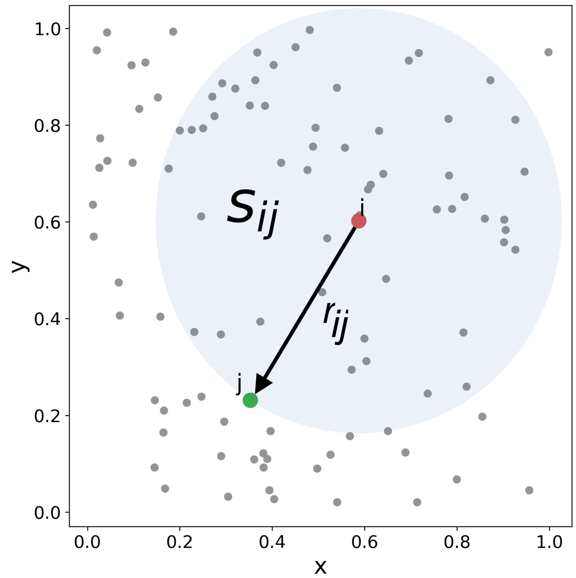


Figure 1.1: Representation of the radiation model. Site i and j are the origin and destination locations respectively and r_{ij} the physical distance between them. The probability of a trip from i to j depends on the total population within the disk s_{ij}.

function, since simple dimensional arguments are not suitable to correctly estimate the value of the distance exponent. An open questions is then how the parameter γ is influenced by changes of the spatial distribution of locations. Moreover, the introduction of the normalisation factor k_i in the equation of the flow breaks the symmetric structure of the unconstrained gravity model so that in general $p_{ij} \neq p_{ji}$.

2.3 The radiation model

The radiation model borrows its main idea from the particle diffusion process in physics. Formulated in terms of job opportunities and job-seeking individuals [17], the law is motivated by the heuristic assumption that a commuting trip takes place when an individual finds the closest job opportunity which offers him/her benefits z higher than the best benefits available in his/her origin location. The model can then be generalised to estimate the volume of trips between two locations based only on the information encoded in the population distribution.

Thus, by considering the probability of an individual travelling between i and j, in analogy with the probability of an absorption event as studied in the physical sciences, the average flux $\langle T_{ij} \rangle$ from location i to location j can be expressed as

$$\langle T_{ij} \rangle = O_i \frac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})}, \tag{1.10}$$

where s_{ij} is the total population residing in a circle of radius r_{ij} (Fig. 1.1), excluding the origin and the destination, and the other terms are as defined in the previous sections (2.1, 2.2).

Moreover, we note that Eq. (1.10) is derived in the thermodynamic limit and therefore requires a normalisation factor for the correct treatment of a finite system. Including this factor, originally derived in [13], we therefore obtain the trip distribution law in the same form as for the gravity model

$$p_{ij} = \frac{O_i}{1 - \frac{m_i}{M}} \frac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})}, \tag{1.11}$$

where O_i is the outflow of location i as defined in Section 2.2 and $M = \sum_i m_i$ is simply the total population in the area of interest.

As can be observed from Eq.(1.11), the estimated flow in the radiation model does not directly depend on the distance r_{ij} and the model, unlike the gravity law, is parameter-free. This represents the most notable feature of the radiation law, which is therefore able to estimate trip volumes only from the population density. Although in the original paper the authors show evidence of scale-invariant properties in commuting flows predicted by the radiation model, we will show later that we do not recover this characteristic under our rescaling framework.

3. Motivation & Outline

Despite the success of these models, dedicated tools for uncovering their scaling behaviour are poorly developed. In essence, in modelling human mobility, both the gravity and the radiation model overlook the spatial dependence of their predictions. The central aim of our study is therefore to comparatively assess the scale-dependence of the gravity and the radiation model and propose a procedure to adapt the gravity law to different coarse-graining levels of a population distribution.

The motivation behind this research is twofold: mobility data might not always be available at the desired resolution; in the absence of a suitable dataset, it is then particularly useful to have a modelling framework that can adapt to a coarse-grained picture of the spatial system. Moreover, as traditional transport surveys become obsolete and small-scale mobile tracking takes its place, multiscale modelling approaches become increasingly important. To address these problems, we explore the following three questions:

1. Does the structural difference in the mathematical formulation of the gravity and the radiation model result in a divergent treatment of different spatial scales? (Chapter 2)

2. What is the error introduced in the predicted trip distribution when rescaling a spatial system in a simplified framework? (Chapter 3)

3. Can we upscale, thus substituting a fine-scale population distribution by a coarser one, without introducing significant error in a real population distribution? (Chapter 4)

In order to provide an answers to the above, we structure our study in the following way:

- We define a quantitative metric (ϵ) of the discrepancy introduced in the estimated flow between locations when coarse-graining a spatial system
- We derive the analytical form of the scaling error ϵ for the gravity (exponential and power-law form) and the radiation model in a simple case (two locations clustered at once within a synthetic population distribution).
- We develop computational simulations of simple spatial systems (random distribution of locations with a uniform population distribution).
- We compare the simulated results to the ones predicted by the theory, finding good agreement between the two, and determine the validity of our assumptions and approximations.
- We extend this procedure to a realistic system by rescaling real UK population distributions in two urban areas by means of the hierarchical agglomerative clustering algorithm.
- We assess the multiscale accuracy of the gravity and the radiation model against our metric and reveal systematic discrepancies between different spatial scales.
- We suggest an optimisation procedure to rescale the distance parameter (exponent γ) in the gravity model so as to minimise the error ϵ.

Chapter 2

Theoretical Analysis

In real mobility flows, the process of aggregating spatial units can be thought of as a linear transformation of the spatial system. For each lower-resolution unit, comprising of n neighbouring units in the original distribution, the resulting collective flow thus equates the sum of the flows attributed to each sub-unit. For example, the number of people moving from London to Birmingham exactly corresponds to the sum of the trips originating from all the boroughs within London and terminating in any area within Birmingham. Therefore, starting at any granularity level, we can always coarse-grain a spatial system (e.g. population distribution) to form bigger spatial units of arbitrary surface areas and recover the real flow between these units. As a result of this, we deduce that the trip distribution per se is invariant or self-similar under a length-scale transformation.

When modelling collective movements of people, however, mobility models introduce highly nonlinear parameters with the aim of reducing the complexity associated with explicitly tracking the trajectory of a person, as done instead in individual mobility models. This makes the rescaling process, and consequently the effect of spatial scale on the models' predictions, a nontrivial problem.

Hence it is an interesting question whether the discrepancies introduced by the models result in near self-similar properties of the mobility network when this is subject to a rescaling procedure or whether the error introduced is in all cases non-negligible. This motivates our theoretical analysis of the problem.

1. Definition of the scaling error ϵ

In order to estimate the error resulting from an arbitrary coarse-graining procedure like the one represented in Fig.2.1, we define the fractional difference between the flow after rescaling the system and the total flow at the base level. The rescaling step simply consists of grouping an arbitrary number of locations into two clusters a and b. The guiding idea is that, since both the radiation and the gravity model produce an output that, directly or indirectly respectively, depends on the population distribution, varying this results in an altered output flow. As the extent of this variation cannot be straightforwardly inferred from the model, it is useful to introduce a new measure to encode this change.

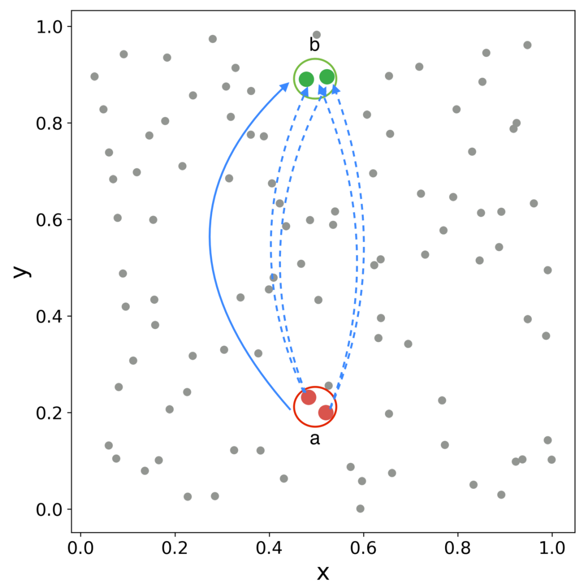


Figure 2.1: Example of an arbitrary aggregation of 4 units into 2 clusters, a and b. The solid arrow represents the flow probability p_{ab} from cluster a to cluster b when the clusters are considered as units within the distribution. The dashed arrow represent the mobility flow probability p_{ij} from each sub-unit i within cluster a to each sub-unit j within cluster b.

The resulting quantity, which we call the *scaling error*, is thus defined as follows

$$\epsilon := 1 - \frac{\sum_{i \in a, j \in b} p_{ij}}{p_{ab}}, \tag{2.1}$$

where p_{ij} represents the number of trips from any location i within cluster a to any location j within cluster b and p_{ab} is the aggregate flow between clusters a and b. The difference between the numerator and denominator in Eq. (2.1) arises from the fact that, after the rescaling step, a and b are treated by the models as single locations. The exact value of ϵ therefore depends on how their position is chosen. In summary, the scaling error ϵ can be regarded as the relative bias introduced by modelling the flows at an aggregated level compared to the spatial distribution at the finest resolution. Since this provides the highest level of information about the variables characterising the spatial system, it is reasonable to assume that, at least on theoretical grounds, the finest resolution level should yield the most accurate representation of mobility patterns for a given system.

2. Toy model of the scaling error

Since the population distribution is typically highly heterogeneous, it is not straightforward to treat our model for the scaling error exactly. In order to be able to deal with the heterogeneity of the population distribution analytically, we first consider a simplified scenario by generating a synthetic population and imposing the following approximations:

- All locations are independently uniformly distributed in space

- The population density is homogeneously distributed among the locations, i.e. $m_i = 1, \quad \forall i$
- Only two locations $\{i, j\}$ are clustered at any one time
- The distance between the two clustered locations $r_{jk} \ll r_{ib}$, the distance between the origin i and the cluster b

For the sake of simplicity, we also restrict our study to forwards flows (from the unclustered location i to the cluster b). Therefore, in this analysis, which we will refer to as *tripoint*, only the parameters of three locations are varied at any one time, while the rest of the population distribution is left unchanged. An estimation of the scaling error is then obtained by probing the difference in the predicted flow when the destination is regarded as two separate sub-areas or as a single zone. Although naive, this method allows us to quantitatively assess the performance of the gravity and the radiation model in terms of their spatial dependence, while abstracting from other effects which may be due, for example, to a heterogeneous population distribution. In this simplified tripoint framework, we obtain the following form for the scaling error:

$$\epsilon = 1 - \frac{p_{ij} + p_{ik}}{p_{ib}}. \tag{2.2}$$

The next two sections provide a detailed derivation of the analytical form of ϵ for the two human mobility models here considered.

2.1 Derivation of the scaling error in the gravity model

Due to the fact that the normalisation factor k_i in the gravity model equation (1.9) depends on the distance between the origin i and all other locations, the form of k_i is affected by the spatial distribution of the sites. Consequently, the predicted volume of spatial interactions varies when changing the granularity level, like done in our tripoint scaling procedure.

The analytical form of the scaling error ϵ for the gravity model with an exponential deterrence function can be derived by substituting (1.2) into (2.2)

$$\epsilon_e = 1 - \frac{m_j e^{-\gamma r_{ij}} + m_k e^{-\gamma r_{ik}}}{m_b e^{-\gamma r_{ib}}} \frac{k_i}{\tilde{k}_i}, \tag{2.3}$$

where we have denoted the normalisation factor after rescaling as $\tilde{k}_i$. We first consider the pre-scaling normalisation factor k_i and decompose it as follows

$$k_i^{-1} = \sum_{l \neq j,k} m_i e^{-\gamma r_{il}} + m_j e^{-\gamma r_{ij}} + m_k e^{-\gamma r_{ik}}. \tag{2.4}$$

In an analogous way, we can then write $\tilde{k}_i$ as

$$\tilde{k}_i^{-1} = \sum_{l \neq b} m_i e^{-\gamma r_{il}} + m_b e^{-\gamma r_{ib}}, \tag{2.5}$$

where we note that

$$\sum_{l\neq j,k} m_i e^{-\gamma r_{il}} = \sum_{l\neq b} m_i e^{-\gamma r_{il}}. \tag{2.6}$$

For a sufficiently high number of locations N, the RHS of Eq. (2.4),(2.5) is dominated by the summation, so that the terms outside it are negligible:

$$\sum_{l\neq b} m_i e^{-\gamma r_{il}} \gg m_b e^{-\gamma r_{ib}} \simeq m_j e^{-\gamma r_{ij}} + m_k e^{-\gamma r_{ik}}. \tag{2.7}$$

As a consequence, $\tilde{k}_i \simeq k_i$ and the expression for ϵ can be reduced to

$$\epsilon_e = 1 - \frac{m_j e^{-\gamma r_{ij}} + m_k e^{-\gamma r_{ik}}}{m_b e^{-\gamma r_{ib}}}. \tag{2.8}$$

Provided that $r_{ib} \gg r_{jk}$, it is then reasonable to approximate $r_{ij} \simeq r_{ik} \simeq r_{ib}$. Since we know by definition that $m_b = m_i + m_j$, we can then assume the approximation in (2.7) holds. In Appendix A we show that this is the case if the total number of locations N is large enough. Finally, noting that

$$r_{ij} \simeq r_{ik} \simeq \sqrt{r_{ib}^2 + \left(\frac{r_{jk}}{2}\right)^2}, \tag{2.9}$$

as clear from the diagram in Figure 3.1, (2.8) can be rewritten as

$$\epsilon_e = 1 - e^{-\gamma(r_{ij} - r_{ib})}. \tag{2.10}$$

The same derivation can be followed for the power law form of the deterrence function $f(r_{ij}) = r_{ij}^{-\gamma}$ to obtain

$$\epsilon_p = 1 - \frac{m_j r_{ij}^{-\gamma} + m_k r_{ik}^{-\gamma}}{m_b r_{ib}^{-\gamma}}. \tag{2.11}$$

For $m_b = m_j + m_k$ and $r_{ij} \approx r_{ik}$, this also simplifies to:

$$\epsilon_p = 1 - \left(\frac{r_{ij}}{r_{ib}}\right)^{-\gamma}. \tag{2.12}$$

Strikingly, in the case of uniform population density, the mass-dependency vanishes and the result for the scaling error is only a function of physical distances. The analytical expressions in Eq.(2.12) and (2.10) reveal that the scaling error decreases with the distance between the origin location and the destination cluster, suggesting, as intuitively expected, that the predicted flow from i is less sensitive to variations in the distribution far away from its position. On the other hand, the equations also tell us that an increasing intra-cluster distance yields a higher scaling error.

2.2 Derivation of the scaling error in the radiation model

Similarly, we can derive the analytical form for the scaling error ϵ when applying the tripoint analysis to the radiation model. In this case, we replace the flow p in the general equation for ϵ (Eq. (2.2) with the radiation trip distribution law in Eq. (1.11) to obtain

$$\begin{aligned}\epsilon =& 1 - \frac{m_i}{\cancel{1 - \frac{m_i}{M}}}\left[\frac{m_i m_j}{(m_i + s_{ij})(m_i + m_j + s_{ij})} + \frac{m_i m_k}{(m_i + s_{ik})(m_i + m_j + s_{ik})}\right] \\ & \frac{\cancel{1 - \frac{m_i}{M}}}{m_i}\frac{(m_i + s_{ib})(m_i + m_b + s_{ib})}{m_i m_b} \\ =& 1 - \frac{\cancel{2}}{(1 + s_{ij})(2 + s_{ij})}\frac{(2 + s_{ib})(3 + s_{ib})}{\cancel{2}},\end{aligned} \tag{2.13}$$

where we have used the fact that $m_i = 1, \forall i$. We can now make use again of the homogeneity of the population distribution to note that, in this case,

$$s_{ij} = \rho\pi r_{ij}^2. \tag{2.14}$$

Similarly to the previous case, if $r_{ij}^2 \simeq r_{ik}^2$, we can further assume $s_{ij} \simeq s_{ik}$ to simplify the expression. We then substitute the equation for s_{ij} and s_{ib} to obtain the final equation

$$\epsilon_r = 1 - \frac{\cancel{\rho\pi} r_{ib}^2}{\cancel{\rho\pi} r_{ij}^2}\frac{(2 + \rho\pi r_{ib}^2)}{(1 + \rho\pi r_{ij}^2}. \tag{2.15}$$

Chapter 3

The Tripoint Model

1. Method: simulation of the tripoint model

In order to validate our model for the scaling error ϵ, we start from the simple tripoint case and simulate the mobility flows by generating a synthetic population with $N \approx 100$ locations randomly distributed on a two-dimensional plane with uniform density. This allows us to easily keep control of all the variables at play and ensure that the variations in the flow are solely due to the rescaling process and not to other effects.

A built-in random number generator is used to produce N pairs of coordinates independently and uniformly distributed across the interval [0, 1], thus effectively populating a unit square through a point process. The synthetic population thus obtained neglects socio-economic and demographic features and purely serves as a proxy for an idealised spatial system. Given our aim of comparing the models on a purely mathematical ground, we are not concerned with reproducing a realistic scenario at this stage.

A tripoint grouping procedure is then implemented on the obtained population distribution with the aim of reproducing the mathematical framework described in Section 2 of Chapter 2. We do so by means of two different approaches and develop two algorithms to implement them, which we call the *sampled tripoint* algorithm and the *explicit tripoint* algorithm. Both these procedures yield the value of ϵ in relation to two quantities of interest: the distance r_{ib} between the origin and the destination cluster b and the distance r_{jk} between the clustered locations (Fig. 3.1). Our ultimate goal is in fact to identify the spatial range over which the error ϵ can be considered small enough so that the models effectively scale with distance, that is, we obtain a near scale-invariant behaviour.

All distance variables in this Chapter are rescaled by their typical separation $\sim \frac{1}{\sqrt(N)}$.

1.1 The sampled tripoint method

The first method to implement the tripoint clustering consists of a probabilistic approach whereby, at each iteration, one pair of nearest-neighbour sites $\{j, k\}$ is grouped together to form a cluster (Fig. 3.1). After this simple clustering step, the value of ϵ is calculated according to Eq. (2.1) and the procedure is iteratively carried out until all possible combinations of origin-destination triads have been explored. The algorithm then outputs the

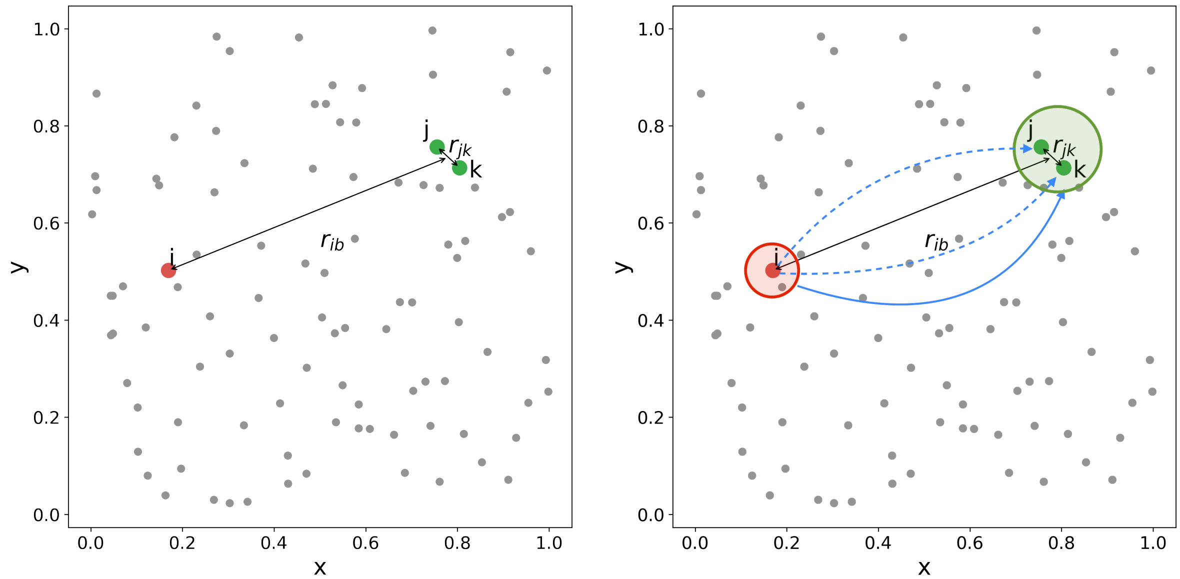


Figure 3.1: Example of a tripoint aggregation. In (a) the origin-destination distance r_{ib} is the Euclidean distance between origin i and the midpoint between j and k. In (b) the dashed arrows represent the flow p_{ij} prior to the clustering step and the solid arrow illustrates the flow probability p_{ib} from i to the new unit b formed by clustering j and k. Similarly, the tripoint configuration comprises, at each iteration, of a cluster made of one among all the possible nearest-neighbour pairs within the distribution.

array of $N(N-1)$ ϵ values resulting from the tripoint clustering of all the possible configurations within the sample. An array of the spatial distances r_{ib} and r_{jk} characterising each configuration is also stored at each realisation of the simulation in order to allow for the scaling error to be analysed as a function of the spatial scale. The nearest-neighbours pairs are found by means of the *ball tree* algorithm implemented in the *scikit-learn* package [1], which was adopted in alternative to a naive linear search to benefit from a reduction of the time complexity from $O(N^2)$ to $O(logN)$. The complete algorithm is outlined in Algorithm 1.

Given the probabilistic nature of the simulation, the synthetic population distributions thus generated differ from one realisation of the simulation to the other and therefore the stochastic behaviour was appropriately taken into account when extracting the results. Specifically, the statistical fluctuations of the the results obtained through the sampled tripoint process were treated with the moving average smoothing technique, in order to expose the underlying distribution. A window of size $w = 50$ was chosen as this was found to sufficiently filter the fluctuations while ensuring that the variance within each window ranged from 0.1% to 3% of the scaling error. However, when dealing with the intra-cluster distance, an additional binning procedure was necessary prior to the smoothing step: since for each nearest-neighbours distance r_{jk}, $(N-2)$ values of the scaling error are found (one for each remaining site within the population), a more accurate result is

[1] http://scikit-learn.org/stable/modules/neighbors.html#ball-tree

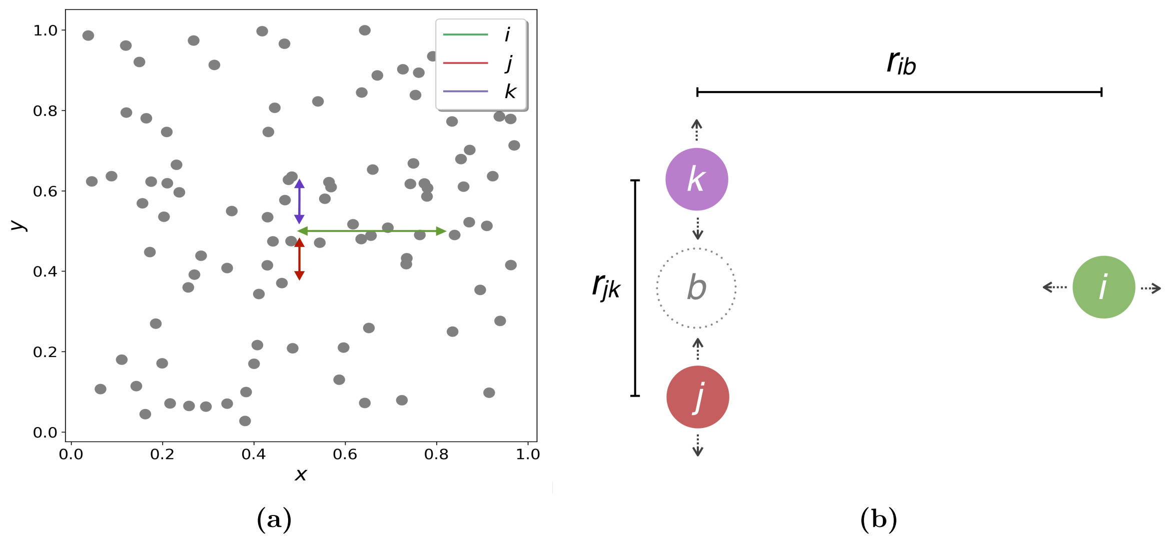


Figure 3.2: Schematic diagram of the explicit tripoint method. At the initialisation step, the origin location i is chosen with coordinates $(x_i, 0.5)$, while the destination units $\{j, k\}$ are taken so that $x_j = x_k = 0.5$. At the clustering step, either r_{ib} $(0.01 < r_{ib} < 5)$ or r_{jk} $(0.01 < r_{jk} < 1.5)$ is varied, by moving the x-coordinate of i or the y-coordinate of j and k respectively, while keeping the distance vectors perpendicular.

obtained by binning the multiple values corresponding to each r_{jk}. From this, the average value and standard deviation were extracted and further processed (Section 2).

1.2 The explicit tripoint method

An equivalent approach, the explicit tripoint algorithm, was developed to concomitantly test the robustness of the tripoint construct by manually varying the position of the clustered locations, as illustrated in Fig.3.2a, instead of sampling pairs of nearest neighbours in the population. In this case, the destination pair {j, k} is horizontally aligned to the centre the unit square $(x_j = x_k = 0.5)$ so that r_{ib} can be easily be varied by simply moving the position of the origin location i. Furthermore, we exploit the statistical independence of the spatial distribution in the point process employed to create the synthetic population in order to obtain reliable results by averaging over multiple simulation runs. This allows us to reduce bias due to sample variability. Unlikely in the sampled tripoint method, in fact, the deterministic variation of the distance between the tripoint configuration (i, j, k) does not provide a . The final ϵ values are then computed by finding the mean and standard deviation over the $n = 500$ runs. To ensure consistency, the same seed was adopted for the random number generator.

1.3 Extracting the tripoint scaling error

As described in Equations (2.10), (2.12) and (2.15), in both the mobility models considered, the scaling error ϵ explicitly depends on two variables: r_{ib}, the distance between the origin location i and the destination cluster b, and r_{jk}, the distance between the two clustered locations. These two distances are varied independently of each other in both the sampled tripoint process (Section 1.1) and the explicit tripoint process (Section 1.2). In the first case, this is guaranteed by the statistical independence of each sampled point in the 2d-space, and in the second by the fact that the two lengths are explicitly varied within a given interval. Therefore, we can observe this double dependence and distinguish the effects of each variable through separate plots of ϵ as a function of r_{ib} and r_{jk} respectively. To do so, in plotting the relationship between the scaling error and the variable of interest, we hold the other quantity constant. We choose these fixed values to be close to the the mean inter-location separation $\langle r_{il} \rangle$ and the mean nearest-neighbour distance $\langle r_{jk} \rangle \simeq 5.2$ [23]. Specifically, for the sampled tripoint method, we can only keep a variable constant when computing the analytical curve, while the numerical results necessarily employ all possible configurations within the system. Moreover, since we restrict the clustered pairs so that they are composed of nearest-neighbours only, we can expect $r_{jk} \ll r_{ib}$ so that our assumptions are reflected in the computational experiment. In contrast, the explicit tripoint method allows us to explicitly fix one of the distance variables and the only constriction in this case is dictated by the finite size of the system.

Algorithm 1 The sampled tripoint algorithm

Input: $\{x_i, y_i\} \quad i = 1, ..., N$ sampled from a uniform distribution
Output: ϵ_{ib}

Generate $P = \{i\}$ // population object
Compute $\mathbf{O}$ // OD matrix
Find the nearest-neighbour pairs
Store their indices $\{j, k\}$ in array A
Store their distances r_{jk} in array R
while $\{x_i, y_i\} \neq \varnothing$ **do**
  Pick i // origin location
  while $A \neq \varnothing$ **do**
    Pick $\{j, k\}$ // destination locations
    if $i \notin \{j, k\}$ **then**
      Create a copy P^{new} of P
      Remove $\{j, k\}$ from P^{new}
      Add b to P^{new} // cluster
      $m_b \leftarrow m_j + m_k,$
      $(x_b, y_b) \leftarrow (\frac{x_i + x_j}{2}, \frac{y_i + y_j}{2})$
      Compute $\mathbf{O^{new}}$ // new OD matrix
      Compute ϵ_{ib} // scaling error
    end if
  end while
end while

Algorithm 2 The explicit tripoint algorithm

Input: $\{x_i, y_i\} \quad i = 1, ..., N$ sampled from a uniform distribution
Output: ϵ_{ib}

1: Generate $P = \{i\}$ // population object
2: Add i, j and k to P
3: Compute $\mathbf{O}$ // OD matrix
4: Create a copy P^{new} of P
5: Add i, b to P^{new}
6: Compute $\mathbf{O^{new}}$ // new OD matrix
7: Compute ϵ_{ib} // scaling error

2. Results: analysis of the tripoint scaling error

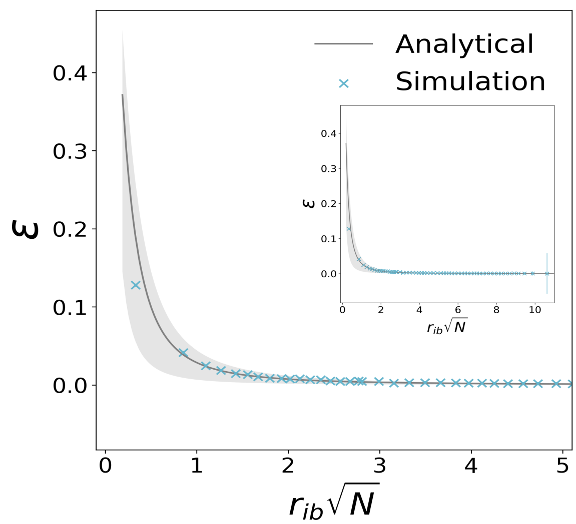
In Chapter 2, we have quantitatively defined the error ϵ introduced in the estimated mobility flow when the resolution of a spatial system is varied. In Section 1 of this chapter we have then constructed and simulated a simplified system, the tripoint configuration, to study the behaviour of the scaling error when only the coordinates of two locations are varied. Here we assess the performance of the gravity and the radiation model against this metric by analysing the numerical results obtained when they are used to generate the mobility flows.

2.1 Tripoint scaling error in the gravity model

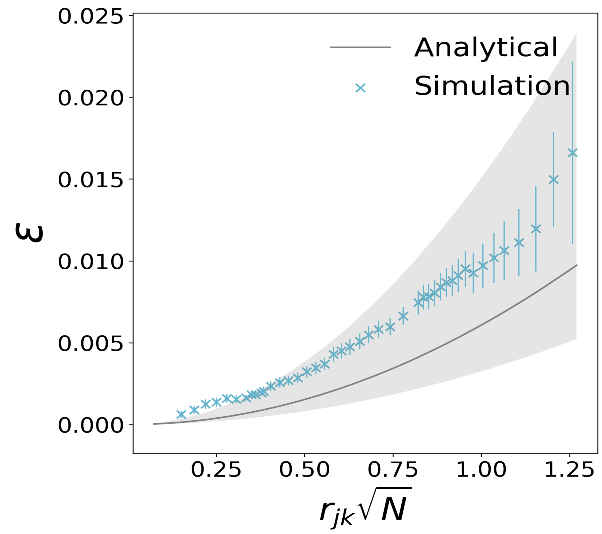
We remark that the gravity model contains an adjustable parameter, the distance exponent γ, which is typically calibrated to the empirical dataset of interest. Since our approach does not involve a comparison with real mobility data, it is not obvious how to estimate this parameter. To this end, we refer to previous results and employ the functional relationship proposed by Lenormand *et al.* (2016), which suggests that, for the power-law form, the distance parameter follows $\gamma_p = 0.3\langle S\rangle^{-0.18}$, while for the exponential form $\gamma_e = 1.4\langle S\rangle^{0.11}$, where $\langle S\rangle$ is the average unit surface (Fig. 4.9) [19]. By considering the typical inter-site separation $\langle r_{ij}\rangle \simeq \frac{1}{\sqrt{N}}$, we can approximate the average unit surface $\langle S\rangle \simeq \frac{1}{N}$, thus obtaining $\gamma_p = 0.84$ and $\gamma_e = 0.69$.

We show here the results obtained for the power-law form of the gravity model. Similar behaviour is exhibited by the scaling error when using the exponential distance decay function and we include the full analysis of this case in Appendix B. All distance quantities are rescaled by the typical inter-site separation $\sim \frac{1}{\sqrt{N}}$.

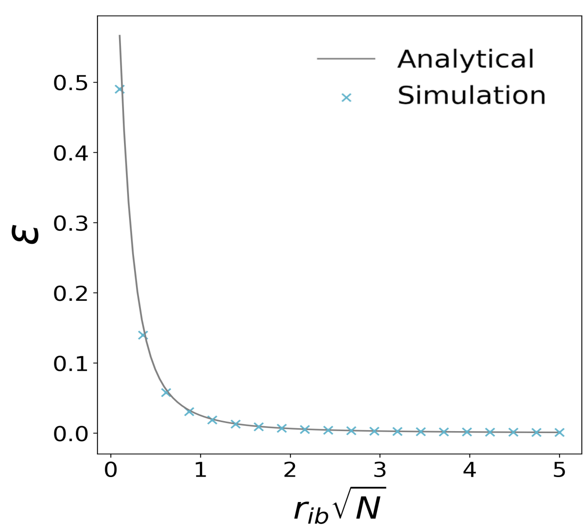
As illustrated in the left column of Figure 3.3, we observe good agreement between the analytical prediction for the scaling error and the data elaborated by our numerical simulation. We find that the value of ϵ exceeds 0.05 only for r_{ib} below the typical inter-site separation, i.e. becomes significant only when the grouping procedure affects the immediate neighbourhood of the origin i. This is expected since the scaling error is based on the idea that the contribution from locations at further distances is small.



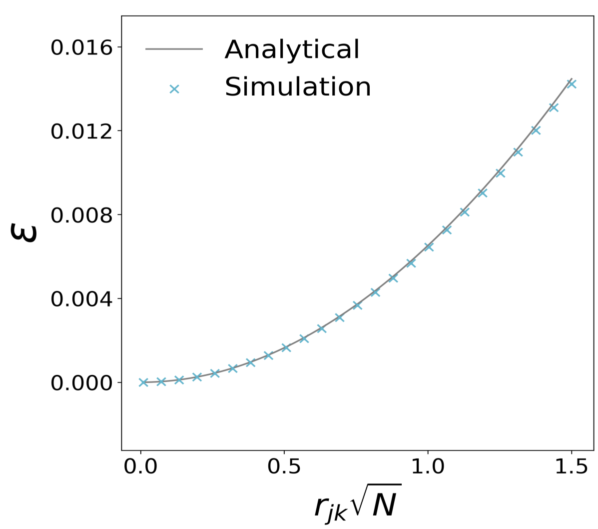
(a) Sampled Tripoint: ϵ vs r_{ib}



(b) Sampled Tripoint: ϵ vs r_{jk}

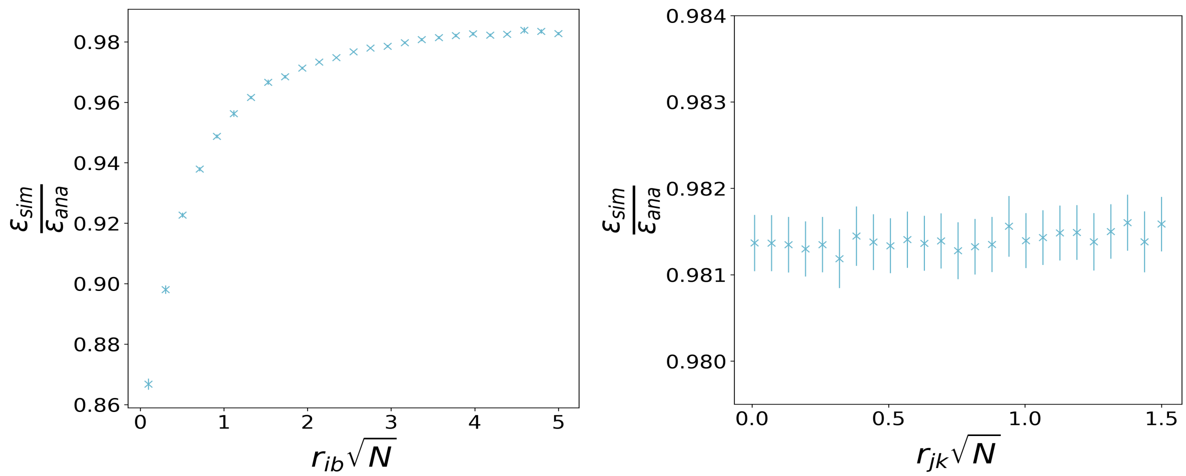


(c) Explicit Tripoint: ϵ vs r_{ib}



(d) Explicit Tripoint: ϵ vs r_{jk}

Figure 3.3: Relationship of the scaling error with distance in the power-law gravity model. The blue markers indicate the numerical results obtained through the sampled tripoint algorithm in (a)-(b) and the explicit tripoint algorithm in (c)-(d), plotted as a function of r_{ib} (left) and r_{jk} (right). The grey curve represent the analytical ϵ for the power-law gravity model in the tripoint aggregation. In all plots, $\gamma = 0.84$ is used as the distance parameter of the deterrence function. In the top panel, $\langle r_{jk} \rangle = 0.5 \pm 0.3$ (a) and $\langle r_{ib} \rangle = 5.2 \pm 2.2$ (b) (extracted from multiple simulation realisations) are used as the fixed values of the distance variable that is not plotted to estimate the value of the analytical result with its confidence interval. In the bottom panel, the fixed distance values are $r_{jk} = 0.5$ (c) and $r_{ib} = 4$ (d), chosen to prevent the results from being affected by edge effects.



(a) **(b)**

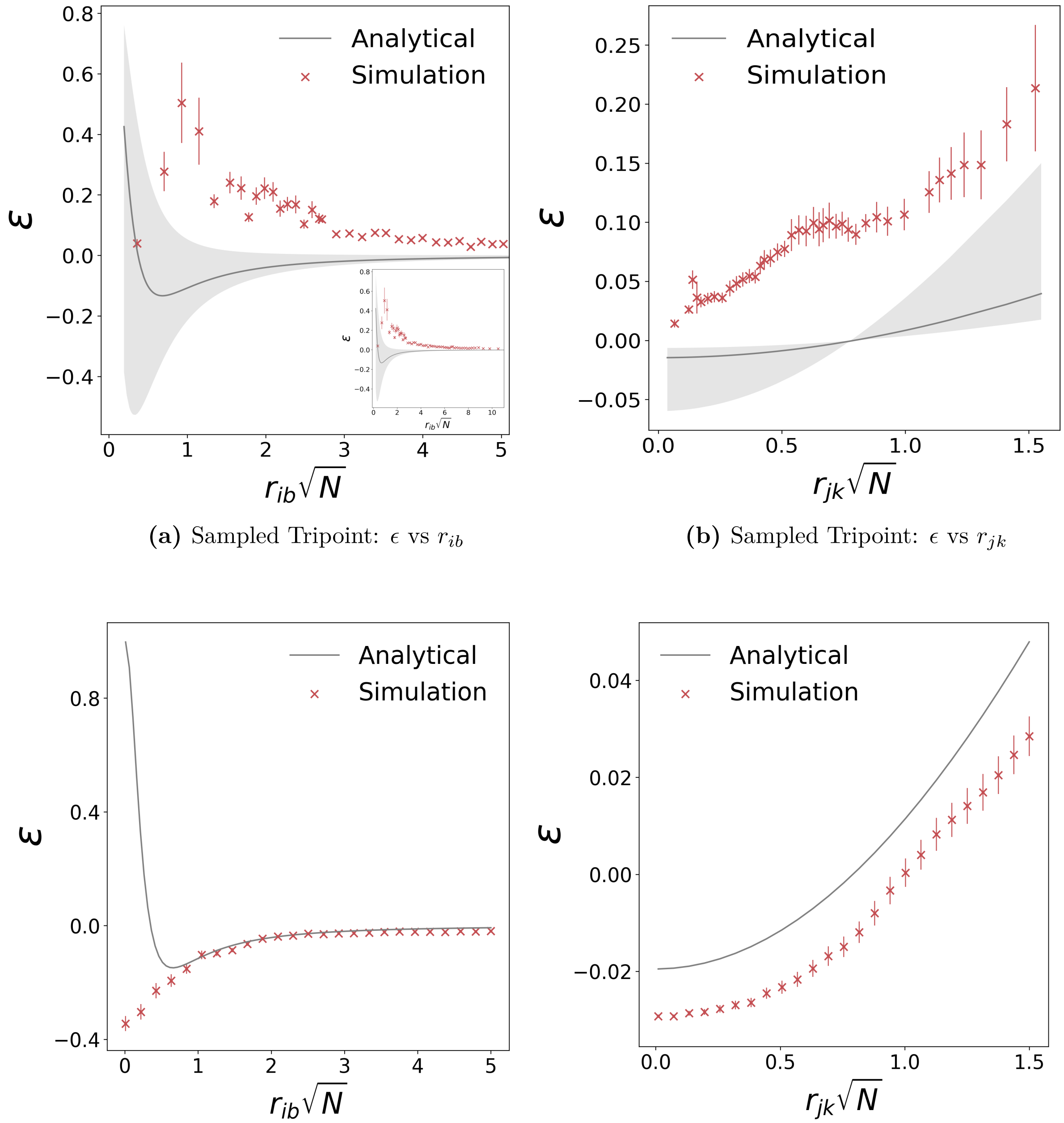
Figure 3.4: Ratio of the scaling error in the simulation and in the theoretical prediction. The ratio does not significantly deviate from 1 in the spatial range studied. However, the degree of agreement with the theory rapidly decreases for $r_{ib} \lesssim 2$. As above, $\gamma = 0.84$ and the fixed distance values are $r_{jk} = 0.5$ (a) and $r_{ib} = 4$ (b).

Conversely, the plots in the right column of Figure 3.3 show that ϵ increases as a function of r_{jk}, consistently with the fact that aggregating units spatially separated by other locations results in considerable differences in the predicted flow. This suggests that sites should only be clustered with elements positioned within their neighbourhood. Nevertheless, even when the cluster b is composed of sites originally separated by a longer distance than the typical inter-site scale, the error ϵ is below 0.02%. We also show the ratio between the predicted and observed scaling error $\frac{\epsilon_{ana}}{\epsilon_{sim}}$ and observe that this does not significantly deviate from 1 in the expected range (Figure 3.4).

2.2 Tripoint scaling error in the radiation model

In the radiation model, we find that tripoint scaling error reaches values above 5% at considerably large origin-destination distances ($5 < r_{ib} < 7$) and small enough intra-cluster lengths ($r_{jk} < 0.25$). This result means that the radiation model is highly sensitive to small variations in the partitioning of the spatial system and suggests that the model lacks the universality property claimed in [17]. This may, at least partially, provide a further explanation for the poor performance of the radiation law in the prediction of intra-urban flows [24].

The bigger discrepancies between prediction and simulation compared to the gravity model arise in this case from the continuous assumption employed in the analytical form of ϵ_r, as Eq. (2.15) makes use of a continuous approximation to estimate the population within the disk s_{ij} around the origin location. While at long ranges this effect is masked by an apparent increase in the homogeneity of the population density, shorter distances make the discrete structure of the synthetic population more prominent. A further limitation caused by the simulated system is also represented by boundary effects which are especially



(a) Sampled Tripoint: ϵ vs r_{ib}

(b) Sampled Tripoint: ϵ vs r_{jk}

(c) Explicit Tripoint: ϵ vs r_{ib}

(d) Explicit Tripoint: ϵ vs r_{jk}

Figure 3.5: Relationship of the scaling error with distance in the radiation model. The blue markers indicate the numerical results obtained through the sampled tripoint algorithm in (a)-(b) and the explicit tripoint algorithm in (c)-(d), plotted as a function of r_{ib} (left) and r_{jk} (right). The grey curve represent the analytical ϵ for the radiation model in the tripoint aggregation. The confidence interval in the analytical curve is estimated by computing ϵ with $\langle r_{jk}\rangle = 0.5 \pm 0.3$ (a) and $\langle r_{ib}\rangle = 4.3 \pm 0.3$, so as to obtain the correct prediction for the simulated configuration. In the bottom panel, the fixed distance values are $r_{jk} = 0.5$ (c) and $r_{ib} = 4$ (d).

relevant for the radiation model, due to the fact that, for locations close to the edges, the total population within the disk is systematically underestimated. However, due to the fractional nature of the measure we study, the bias introduced by this effect is minimised.

We further note that we find a systematic offset of the analytical curve from the simulated values, which is not reflected between the two tripoint approaches. We attribute this to the asymmetries between r_{ij} and r_{ik} that characterise the sampled tripoint configuration, in which nearest-neighbours pairs are not exactly equidistant from the origin location i. In contrast, this symmetry is held in the explicit tripoint method.

3. Discussion

The proposed method successfully reproduces, within the limitations due to finite-size effects and the discretisation of space, our theoretical predictions and extracts the key variation in the estimated flows when two locations are clustered together in a uniform distribution of locations with homogeneous population density. In agreement with our analytical results, we find that the discrepancies between predicted flows at the original resolution level and after clustering two sites vanish in the expected distance range. This spatial range is found to reflect the heuristic observations that:

- the intra-cluster distance r_{jk} must be sufficiently small so as to prevent nearby sites from altering the estimation
- the origin-destination distance r_{ib} should be large enough so that the small variations in the location distribution due to the clustering do not produce significant differences

Accordingly, we conclude that we can identify an optimal range wherein the scaling error ϵ can both be accurately predicted and is sufficiently small, which we call the *scaling regime*.

No significant discrepancy is observed between the results produced using the two different tripoint approaches, the sampled tripoint and the explicit one, a clear indication of their equivalence. The fluctuations observed in the sampled tripoint (Fig. 3.3a, 3.3b) are an expected effect of the stochastic nature of the approach. Moreover, in this case, the degree of agreement with the analytical result also decreases more rapidly at small origin-destination scales (Fig. 3.3a), since the difference between r_{ij} and r_{ik} ceases to be negligible, in contrast to what is assumed in the derivation.

With regards to the performance of the gravity and the radiation model, we observe that in the latter the error rises considerably quicker, suggesting the model adapts less to small changes in the population distribution. This can be easily understood in terms of the mathematical formulation of the model, which implies that the trip probability does not directly depend on metrical distance and therefore its variation with the spatial scale is particularly nontrivial. This is also consistent with findings presented in exiting literature [13, 25] and highlights that the parameter-free property of the radiation model may represent a drawback with respect to the model's reliability in a multiscale context.

Chapter 4

Hierarchical Spatial Scaling

1. Methodology motivation

Thus far we have considered a toy model that allowed us to quantitatively explore the correlation between changes in the predicted flow and the distance measures of interest within a square lattice populated by randomly sampled locations with uniform population density. In order to evaluate to what extent the gravity and the radiation model can adapt to varying spatial scales in a non-uniform population distribution, we now extend our multiscale analysis to a more realistic scenario. Using our scaling error model, we attempt to capture the change in the predicted mobility flow when the system undergoes a complex renormalisation process which changes the spatial scale over which the interactions occur. This also allows us to determine to what extent our scaling error model is consistent with a more generic spatial system. In this respect, the simple tripoint procedure, although useful in capturing some of the fundamental effects related to the models' scale-dependence, cannot suffice, due to its oversimplification of some important spatial features that characterise real population distributions. Therefore, building upon our first method, we relax the approximations previously imposed (see Chapter 2, Section 2) and consider a non-uniform distribution of locations with inhomogenous population density. We then iteratively carry out a coarse-graining procedure to generate a hierarchy of population distribution levels with different granularity.

Thanks to the availability of spatial data at a sufficiently high resolution, we choose to conduct our analysis on the real population distribution of England, Wales and Scotland. However, despite the clear advantages of employing a real population distribution, this presents several challenges from a mathematical and computational viewpoint. Firstly, due to the heterogeneity of the distribution, we can no longer simplify the problem and identify an analytical expression for our scaling error ϵ. Therefore, we approach this by analysing numerical results only. Secondly, handling a dataset of over 41,000 locations significantly increases the computational cost of the simulation. Hence, on the ground of computational ease, we restrict our analysis to subsections of the entire dataset, with a particular emphasis on urban areas.

2. The population dataset

The dataset used for our analysis of a real spatial system is the CDRC 2011 Population Weighted Centroids (LSOA/Data Zone) dataset [26]. This socio-geographical dataset comprises of population-weighted centroids of the Census Output Areas of England, Wales and Scotland obtained from the most recent census, undertaken in March 2011. The partitioning, which has a resolution of 20 m, follows the Lower Layer Super Output Areas (LSOA), regions obtained by aggregating adjacent areas at the lowest geographical level produced by census estimates, the Census Output Areas (OA). LSOAs were introduced with the aim of improving small area statistics [27] and are defined so as to comprise a population ranging from 1,000 to 3,000 people and a number of households varying between 200 and 1,200. The dataset features a total number of 41,729 spatial units, each associated with the Eastings/Northings coordinates of its population-weighted centroid and the total population residing within the area.

2.1 Data preprocessing

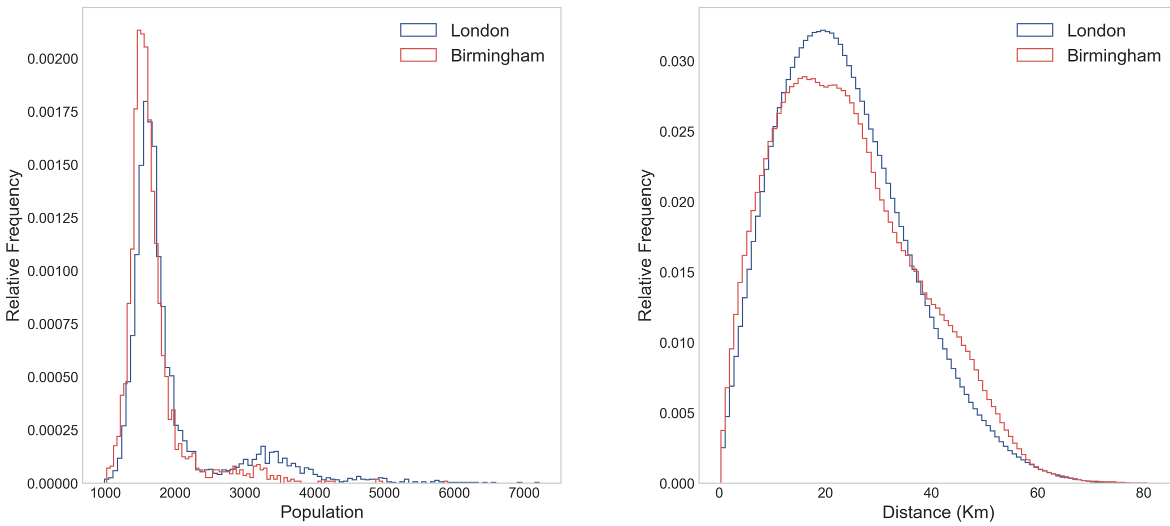
We choose two representative urban areas within our dataset, Greater London and Birmingham, to perform our multiscale analysis. The choice to focus on the two cities with the highest population in the UK is motivated by both practical and scientific reasons: on the one hand this allows us to perform faster simulations and analyses, on the other it ensures that the coarse-graining procedure can be applied to long enough distances without incurring in boundary problems such as edge effects or the presence of natural barriers, which may introduce discontinuities in human interactions [28, 29]. We set artificial boundaries around the urban areas so as to include all the main road networks and extract from the dataset the corresponding units by searching for the Eastings/Northings coordinate pairs of the population-weighted centroids falling within the boundaries.

For the purposes of our analysis, we require two types of information from this dataset, the pairwise distance r_{ij} between every site, and the total population m_i residing within each unit. We find the first by computing the distance matrix $\mathbf{R}$, taking care to employ the condensed form of the matrix for maximum computational efficiency [1]. We note that, since the coordinate system used is based on the Transverse Mercator projection, the distance between two points does not generally correspond to the distance measured on the surface of the earth, but we can safely neglect this effect since in our case the distortion does not exceed 0.01% [30]. A further preprocessing step was carried out to merge the data with an additional dataset containing information relative to the surface area covered by each unit [31], necessary for the purpose of the parameter

City	Number of Units	Surface Area (km^2)	Population Size
London	6,061	3,852	10,172,889
Birmingham	2,415	3,777	3,906,168

Table 4.1: Dataset Statistics

[1] We use the `pdist` function from the SciPy package
https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.pdist.html



(a) Population distribution

(b) Distance Distribution

Figure 4.1: Frequency distribution of population size and inter-unit distance. For both the datasets relative to London and Birmingham, the relative frequency of the population size displays a similar distribution (a) and the pairwise distance (b) can be fitted by a Gaussian. Given their consistency, a comparison between the two areas can be carried out.

estimation in the gravity model. Following the extraction of the data points of interest, we then use these to construct a fine-grained population distribution which corresponds to the first resolution level of the hierarchy.

We present an overview of the areas analysed in Table 4.1. The population distribution and spatial distance distribution at the finest level were plotted for both cities to ensure consistency (Fig. 4.1).

3. Method: rescaling procedure

3.1 Hierarchical agglomerative clustering

As previously stated, we seek to coarse-grain the population distribution so as to construct a hierarchical system composed of decreasing granularity levels. Analogously to the process adopted for the tripoint analysis (Chapter 3, Section 1), this then allows us to compute the scaling error ϵ resulting from variations in the distribution at each of these levels. To this end, we require a renormalisation procedure that iteratively rescales the system, i.e. that groups sets of neighbouring locations $\{i\}$ together and replaces them with a single site $\mathcal{A}$ of population $m_{\mathcal{A}} = \sum_i m_i$. In other words, our aim is to produce a set of levels where the average distance between neighbouring locations progressively increases, thus allowing us to "zoom out" and compute the mobility flows on a coarse-

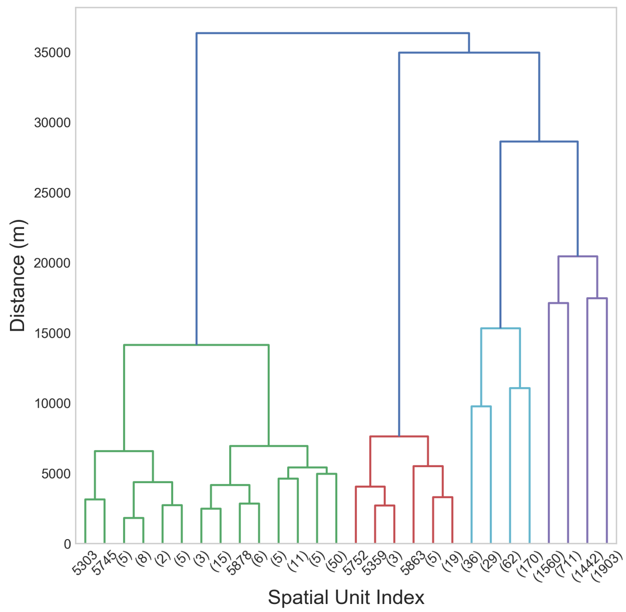


Figure 4.2: Agglomerative hierarchical clustering dendrogram representing the 4 highest clustering levels for the London dataset. The y-axis indicates the pairwise Euclidean distance between the clusters, used as the dissimilarity index between them. The 28 original units are grouped with their closest location at the first stage and the process is iteratively carried out until all units are grouped belong to one agglomeration. As a result, the effective distance between the locations increases at each step. This mechanism allows us to define our hierarchy by choosing appropriate maximum clustering distance values (d_{max}).

grained picture of the spatial system. In order to distinguish between two levels, we label units at the pre-clustering stage with $\{i, j, k...\}$ and use $\{\mathcal{A}, \mathcal{B}, \mathcal{C}...\}$ to denote the clusters.

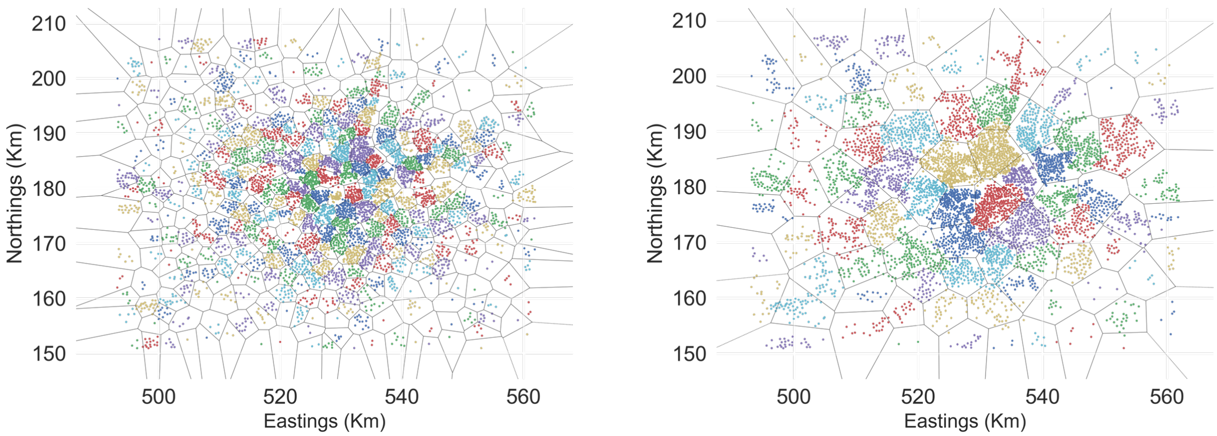
To accomplish this, we make use of the *agglomerative hierarchical clustering* algorithm [32], which we apply to our fine-grained population distribution. This is an efficient unsupervised machine learning technique [2] that produces a dendrogram of the input data (Fig. 4.2), in which each level is built from the previous on the base of a dissimilarity index, here chosen to be the Euclidean distance between locations.

The algorithm takes as input the $N \times N$ *distance matrix* $\mathbf{R}$

$$\mathbf{R} = \begin{bmatrix} 0 & r_{12} & \dots & r_{1N} \\ r_{21} & 0 & \dots & r_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ r_{N1} & r_{N2} & \dots & 0 \end{bmatrix}, \tag{4.1}$$

where the ij^{th} entry is r_{ij}, the pairwise distance between element i and j, and N is the total number of elements in the set (spatial units in a geographical area, in our case). The set is partitioned initially into singleton clusters, i.e. subsets containing only one element,

[2] We use the SciPy implementation of the algorithm, which has $O(n^2)$ time complexity [33].



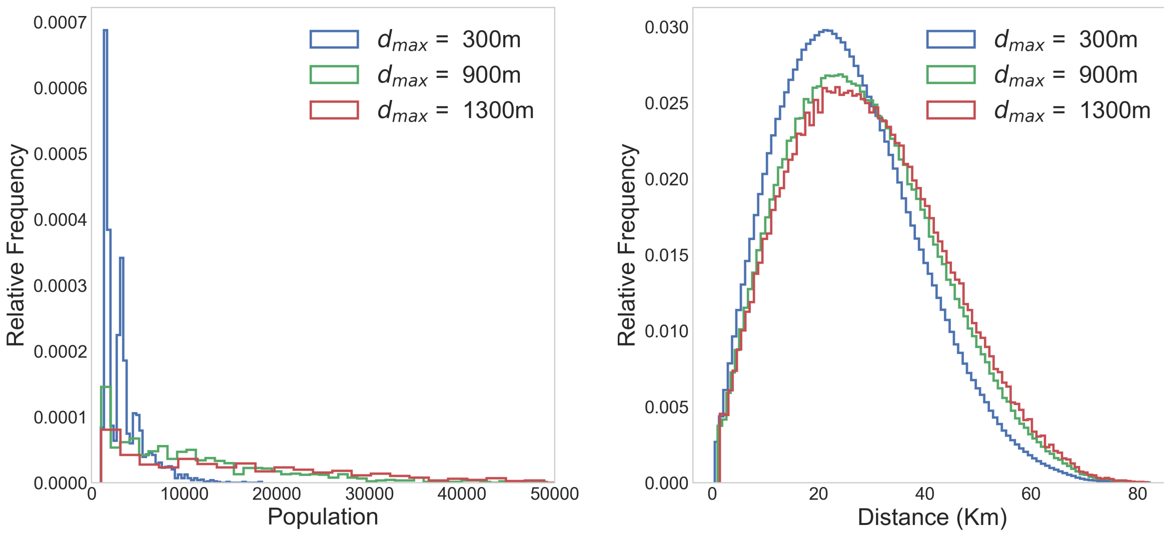
(a) **(b)**

Figure 4.3: Voronoi tessellation of London clusters at two coarse-graining levels, $d_{max} = 2000$ m (a) and $d_{max} = 5000$ m. As the resolution level is decreased, the average unit surface and average inter-site distance increases. Therefore, by applying the mobility models on the same population distribution at different clustering levels enables a comparison of the models' predictions across diverse spatial scales.

and subsets are subsequently joined together into larger clusters in a hierarchical fashion, by merging, at each step, pairs of elements with the smaller distance between them. Therefore, at the final step, the algorithm always returns one cluster containing all the N initial elements. Since we are not interested in such a coarse partitioning of our spatial data, at each level, we introduce a threshold, d_{max}, which restricts the maximum distance between any two neighbouring clusters. Once this distance is reached, the algorithm stops and returns the sets of clusters forming a given level. The use of the threshold therefore enables us to have better control over the spatial scale characterising each level. We note that a key difference with the previous tripoint method is that geographical locations are here not points but regions that extend over a surface area. Therefore, their *position* is not uniquely defined and choosing how to assign spatial coordinates to a new cluster without introducing bias in the statistical distribution is a particularly puzzling problem. This issue is known in the literature as the *modifiable areal unit problem* [34] and, despite being known to researchers for a long time, does not have a definite solution. Our approach to overcome this problem is here to make use of the *centroid* method [35], whereby at each iteration, the distance matrix $\mathbf{R}$ is updated with the distance between the centroids $c_{\mathcal{A}}$ and $c_{\mathcal{B}}$ of all pairs of clusters $\{\mathcal{A}, \mathcal{B}\}$ so that their pairwise distance becomes

$$r_{ab} = ||c_a - c_b||_2, \tag{4.2}$$

where $\mathcal{A} = \{i\}$ is the set of units aggregated into cluster $\mathcal{A}$ and $\mathcal{B} = \{j\}$ is the set of units aggregated into cluster $\mathcal{B}$. Moreover, the choice of employing the centroids as the position of aggregations of spatial units reflects a common practice in the field of spatial analysis, thanks to the fact that this is often found to be sufficiently consistent with empirical



(a) Population Distribution

(b) Distance distribution

Figure 4.4: Frequency distribution of population size and inter-unit distance at 3 of the clustering levels used in our study for the London dataset. While the coarse-graining process preserves the overall distance distribution and only affects its mean value and spread, the population distribution is effectively made more homogeneous compared to the original partitioning.

observations of mobility flows [19, 36]. Although the population-weighted centroid is used in the original dataset instead, the clustering procedure here used effectively smooths out the population distribution within a cluster (i.e. this is uniformly distributed across the entire surface area), so that the population-weighted centroid corresponds to the centroid itself. An example of the resulting partitioning resulting from carrying out the clustering procedure here described is illustrated in Figure 4.3 for the London region.

The choice of a clustering algorithm is not unique and a number of alternative algorithms, like the K-means, have also been used to group administrative units in the context of human mobility [16]. However, these methods often do not allow for an explicit use of distance as a parameter in the clustering and instead provide control over the total number of clusters produced or the number of elements within each cluster. Since we are not concerned with either of these two measures directly, the hierarchical clustering we make use of presents clear advantages, since in our case the spatial scale is the central parameter of the analysis.

3.2 Extracting the hierarchical scaling error

The population distribution thus defined at each level can be used to simulate the mobility flow between clusters according to both the gravity and the radiation model and infer how the predicted flow is affected by the change in spatial scale by computing the corresponding scaling error ϵ. Unlike in the simple tripoint case, the scaling error may now be defined relative to multiple granularity levels and therefore we employ the original population distribution (base level) as a unique reference point for the hierarchy.

The stages of the complete renormalisation scheme are outlined below:

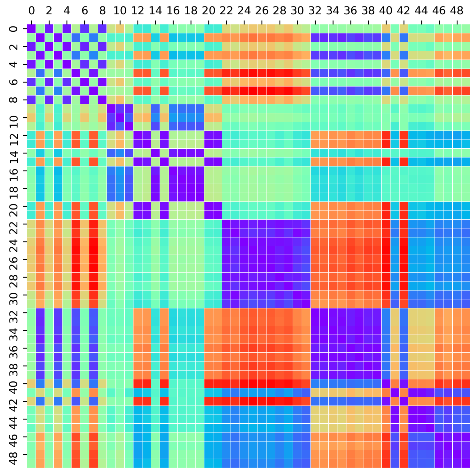
1. Compute the original OD matrix $\mathbf{T_0}$ of the fine-grained population distribution
2. Define n resolution levels by choosing a set of values for the maximum inter-cluster distance $\{d^1_{max}, d^2_{max}, ..., d^n_{max}\}$
3. Apply the hierarchical agglomerative clustering to form the n lower-resolution levels
4. Compute the OD matrix $\mathbf{T_n}$ at each n level by running the mobility models

In order to verify that our algorithm yields the correct output, simple tests were performed by applying the clustering routine to small sections of the dataset comprising of $\sim$ 100 locations and analysing the resulting classification of units within clusters. Tests were then carried out on the regions of interests, London and Birmingham, and the resulting frequency distribution of inter-site distance and population size was plotted to verify Several possible values for the threshold distance were also tested to determine our hierarchical population distribution and found that the range 300m $< d_{max} <$ 1300m, with an interval of 200m provides a suitable set of resolution levels (Fig. 4.4).

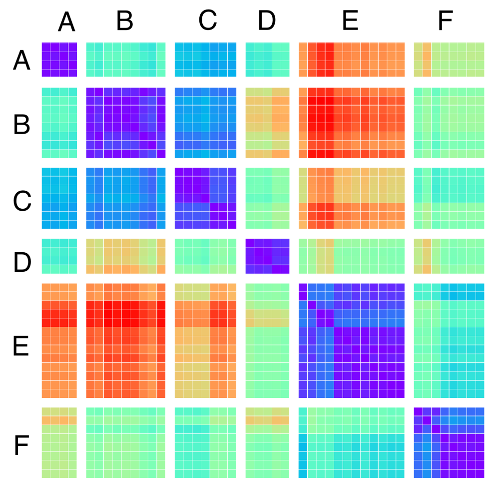
Design of efficient matrix computations

Since this approach involves aggregating an arbitrary number of units at each step and we are interested in extracting a meaningful measure of the scaling error in the system, it is useful to define the error matrix E, where each entry $\epsilon_{\mathcal{AB}}$ represents the fractional difference in the flow from cluster $\mathcal{A}$ to cluster $\mathcal{B}$ compared to the sum of the flows from $\{i\}, \forall i \in \mathcal{A}$ to $\{j\}, \forall j \in \mathcal{B}$, as defined in Eq. (2.1). These poses two challenges: one is that of determining how to accurately extract from the matrix E a measure that best encapsulates the error introduced in the whole system. A second challenge regards how to efficiently implement the necessary matrix operations.

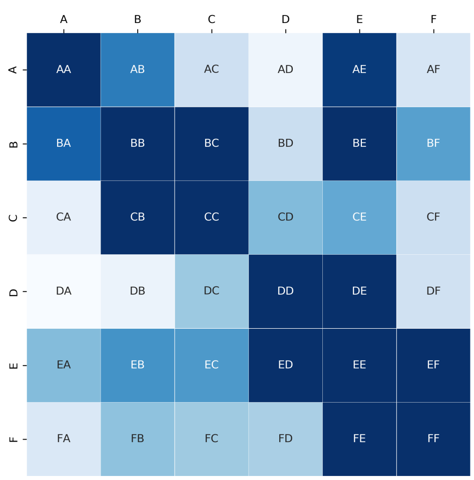
In order to overcome the two aforementioned challenges, we designed an algorithm that performs a renormalisation scheme on the E matrix. The technique makes use of efficient vectorised code to compute E at each level. In fact, particular care is required in carrying out this step, since, in order to subtract the mobility flow at a coarse-grained level from the flow at the base level, the dimensionality of the origin-destination (OD) matrix at the higher-resolution level needs to be reduced to match that of the OD matrix at the lower-resolution level. For example, if the pre-clustering distribution features N locations and, following the clustering step, the number of units is reduced to $M < N$, the error matrix E will have dimension $M \times M$ and hence can only result from the (negative) addition of 2 OD matrices of equal dimensionality. To this end, we develop a dimensionality reduction technique, consisting of a reordering step and a summation step, as described below and illustrated in Figure 4.5. The key advantage of this method is that the scaling error matrix can then be easily computed from the straightforward matrix subtraction $\mathbf{T'_n} - \mathbf{T_n}$. Several variations of this algorithm were tested before finding this was the most computationally efficient.



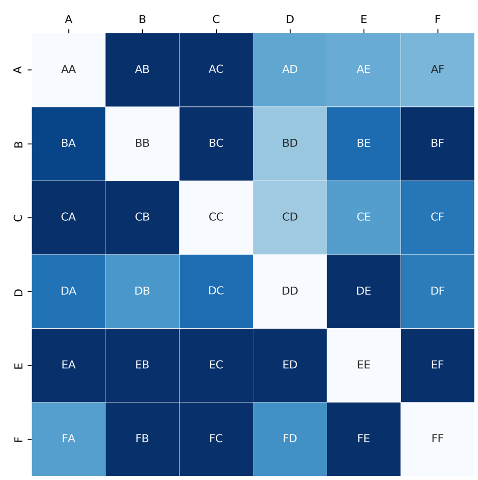
(a) original DM



(b) reordered DM



(c) combined ODM



(d) reduced ODM

Figure 4.5: Renormalisation procedure of the OD matrix. (a) The base-level distance matrix (DM) is computed on the higher-resolution population distribution. (b) Entries in the DM are reordered according to the labels obtained from the cluster classification so that units grouped within the same cluster correspond to adjacent rows and columns. (c) All the flows within the quadrants thus found are summed to obtain the resulting $M \times M$ *combined* OD matrix. (d) The $M \times M$ *reduced* OD matrix $\mathbf{T}'$ is obtained by running the mobility model on the new clusters distribution. Here, the radiation model is used to compute the ODMs.

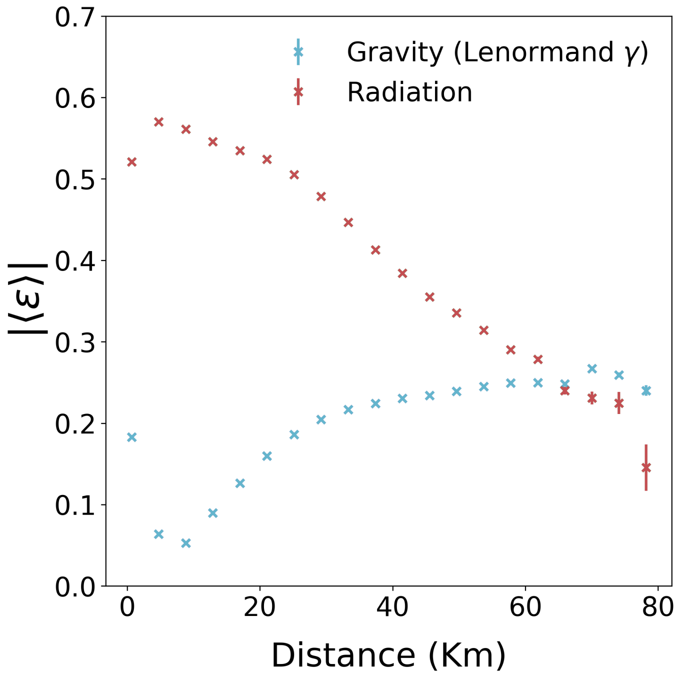


Figure 4.6: Absolute value of the scaling error in the gravity (blue) and radiation (red) model at the a coarse-graining level with $d_{max} = 700$ m. In the range considered, the radiation model generally performs worse than the power-law gravity model, yielding a significantly bigger error. From $r_{\mathcal{AB}} > 60$ km, however, the two models yield comparable results.

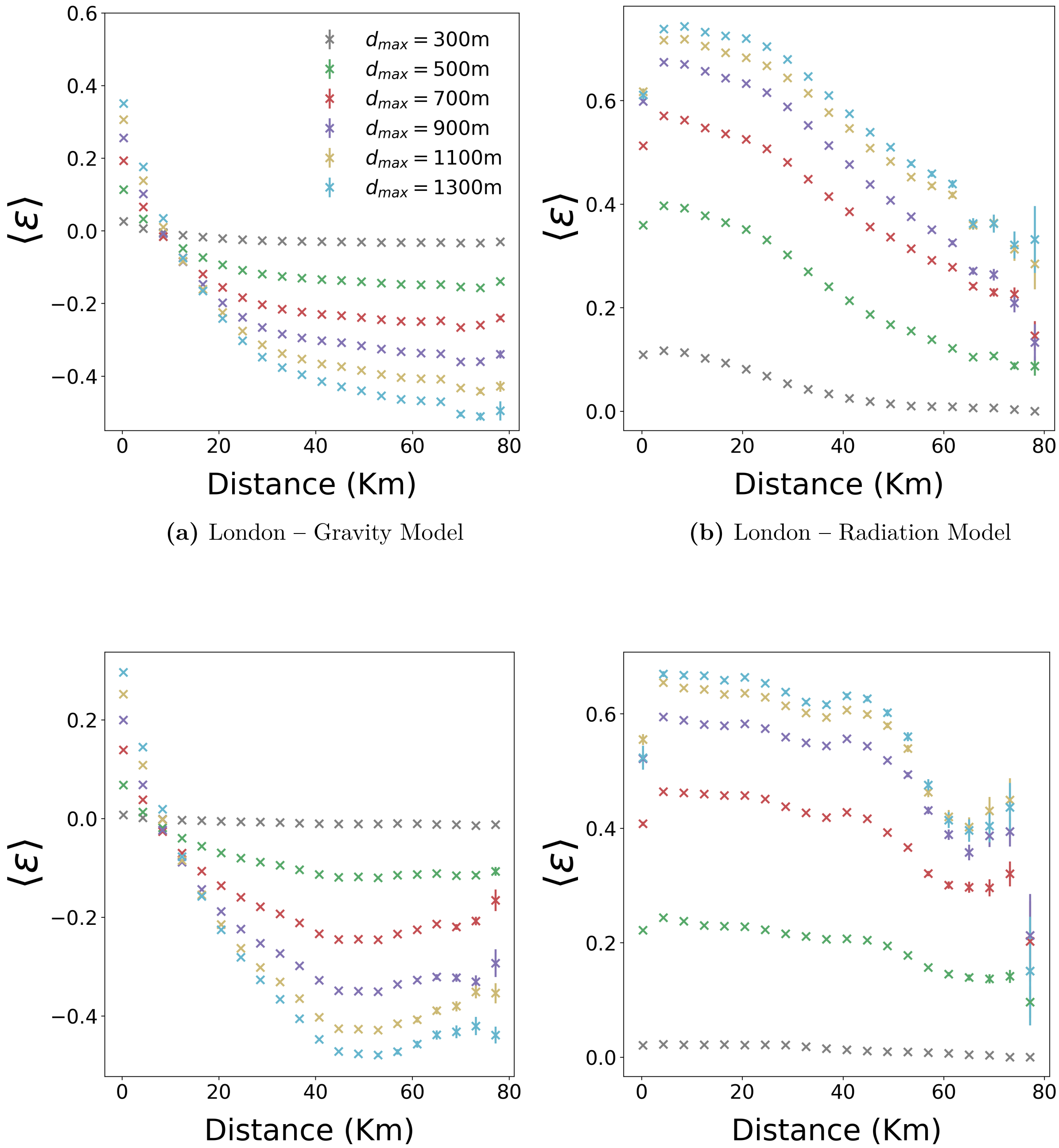
4. Results & Discussion: the hierarchical scaling error

In this section, in analogy with the analysis carried out for the tripoint clustering method, we highlight the main results obtained through the analysis of the scaling error matrix E as a function of the inter-cluster distance $r_{\mathcal{AB}}$, which we will call *distance*.

4.1 Performance of the gravity and the radiation model

Figure 4.7 displays the mean value extracted from the E matrix at each coarse-graining level for the population distribution of London (top) and Birmingham (bottom) across the 6 levels composing the hierarchy in the power-law gravity model (left) and the radiation model (right). As previously done for the tripoint analysis, we employ Lenormand's functional relationship to obtain an estimate of the distance parameter γ.

In order to provide a clear parallel analysis of the two models, we distinguish two separate scaling regimes, one characterising the distance range $0 < r \lesssim 10$ km, which we will refer to as the *short-range* regime, and one characterising the distance range $r \gtrsim 20$ km, the *long-range* regime. Within the small range regime, both models, as expected, produce a higher scaling error at low inter-cluster distances. Although this can be attributed in both cases to an increased apparent homogeneity of the location distributions when estimating flows at higher distances, it results from different mechanisms in the two models.



(a) London – Gravity Model

(b) London – Radiation Model

(c) Birmingham – Gravity Model

(d) Birmingham – Radiation Model

Figure 4.7: Relationship of the scaling error with inter-cluster distance in the power-law gravity model (left) and the radiation model (right) for London (top) and Birmingham (bottom). The clustering levels, plotted each in a different colour, correspond to a maximum inter-cluster distance in the range 300 m $< d_{max} <$ 1300 m. The absolute value of ϵ increases with higher clustering level, since the differences introduced in the popoulation distribution become more significant as the system is upscaled. The standard errors resulting from the data binning process are shown through the error bars, but are in most cases smaller than the data point.

For the gravity law, since the flow is directly dependent on the distance, variations in the spatial structure at a sufficiently long distance will not significantly affect the predictions. In the case of the radiation model, instead, a heuristic explanation may come from considering the disk s_{ij} (Fig. 1.1): if we compare the relative difference produced by small variations of the spatial distribution within a disk of small radius with that produced from variations within a considerably bigger one, it is clear that a smaller scaling error will result in the latter. Hence, at this distance range, the main difference between the gravity and the radiation model resides in the magnitude of the scaling error.

In the long range regime, however, the two models exhibit different behaviour. While the radiation model yields a scaling error which clearly decreases with the inter-cluster distance, this is reflected in the gravity model only for high resolution levels. This is a clear indication of the inadequacy of a unique value of the distance parameter γ to provide consistent predictions across different resolution levels, even for small changes in the spatial scale. When aggregating zones to a resolution of 700 m or lower (higher $\mathrm{d}_{\mathrm{max}}$), as done in our 3$^{\mathrm{rd}}$ clustering level, the model underestimates the mobility flow compared to the base level. This behaviour is more clearly shown in Figure 4.6, where the absolute value of ϵ is plotted for the gravity (blue) and radiation model (red) at the 3$^{\mathrm{rd}}$ clustering level. However, despite this, the gravity model outperforms the radiation model, confirming that the latter cannot offer satisfactory multiscale predictions [37].

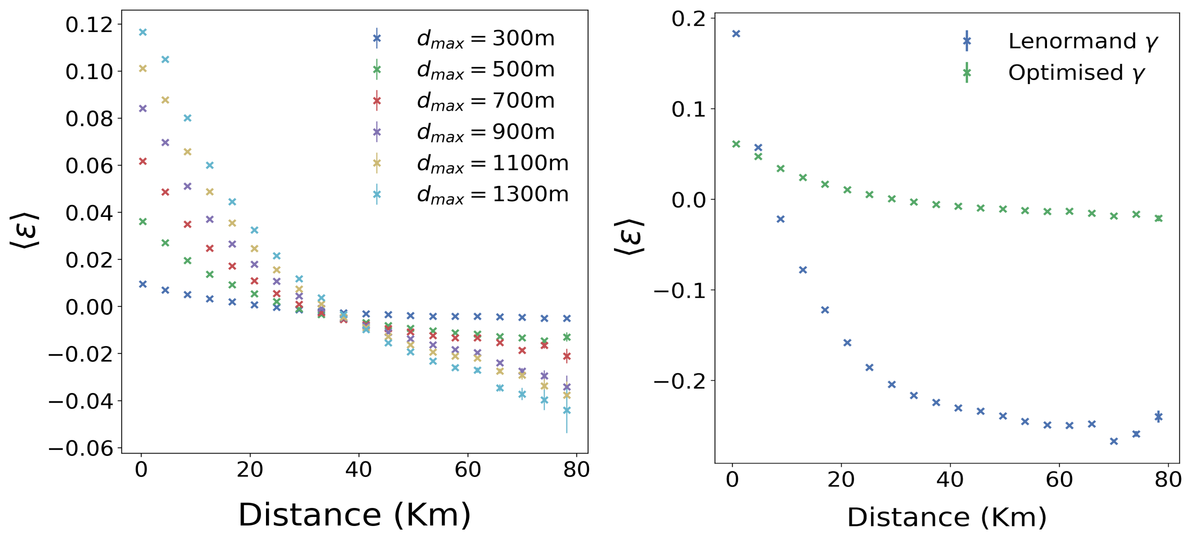
5. Method: optimisation procedure for the gravity model

We now explore a simple approach aimed at tuning the deterrence function in the gravity model so as to minimise the error introduced when changing coarse-graining level. In other words, we devise a procedure to render the gravity model more robust to changes in the spatial scale, without introducing additional parameters. As the deterrence function and its distance parameter γ fully characterise the spatial component in the model, we can probe the parameter space to extract the value that optimises the scaling error ϵ. Although a number of studies suggest a relationship between the distance exponent γ and the fractal dimension of the system [20, 38], this avenue has not been proven successful. Instead, we adopt a different perspective and attempt to treat this as an optimisation problem.

We define the following objective function

$$Z(\gamma) = \sqrt{\sum_{\mathcal{AB}} |\epsilon_{\mathcal{AB}}(\gamma)|^2}, \tag{4.3}$$

where the sum is over all the possible clusters pairs within the configuration at a given coarse-graining level and $\epsilon_{\mathcal{AB}}(\gamma)$ the corresponding entry in the E matrix. As an efficient computational technique to identify the optimising parameter, we employ the SciPy package implementation of the quasi-Newton algorithm [39]. In order to ensure that the optimisation of the function $Z(\gamma)$ leads to a meaningful parameter estimation, we test its convexity of the objective function by plotting it as a function of the parameter and find that it is sufficiently well-behaved. Given the time constraint, we implement this method only for the London population distribution and the power-law gravity model.



(a) (b)

Figure 4.8: Scaling error in the optimised gravity model on the London dataset. The error is computed relative to the flow estimated through the power-law gravity model on the base-level distribution with parameter $\gamma = 1.33$, estimated based on the empirical relationship in [19]. For comparison, we plot in (b) the scaling error at level 3 (maximum inter-cluster distance $d_{max} = 700$ m) in the power-law gravity model with γ computed through Lenormand's empirical relationship (blue) and through our optimisation method (green). A significant reduction in the error is observed through our approach.

6. Results & Discussion: the optimised gravity model

6.1 Performance of the gravity model with rescaled parameter γ

The effect described in Section 4.1 is an intrinsic property of the gravity model and our results provide clear evidence that the choice of resolution level significantly affects the fitting of the distance parameter. This has important implications in the calibration of the model's parameters and can explain the need to often employ different exponents even within the same dataset. Nevertheless, since this is a structural property of the gravity model, the effect cannot be completely eliminated without introducing further tunable parameters. We disregard this solution on the grounds that multiple parameters are an undesirable feature, especially in data-scarce contexts. However, our results suggest that, provided that a sufficiently accurate base level can be employed to estimate the spatial interaction at a fine resolution, the deterrence function can be adjusted so as to minimise the discrepancy in the estimated flow of a lower resolution level. Although this finding would require further validation through fitting of empirical flows, we provide a preliminary validation of this hypothesis by examining the mean scaling error across the whole system as a function of the inter-cluster distance when our rescaled parameter $\tilde{\gamma}$ is used in alternative to the one proposed by Lenormand *et al.* (2016). In figure 4.8 we show that our method provides a notable improvement in the multiscale performance of the power-law gravity model, with a reduction of the scaling error both in the short-range and in the long-range regime.

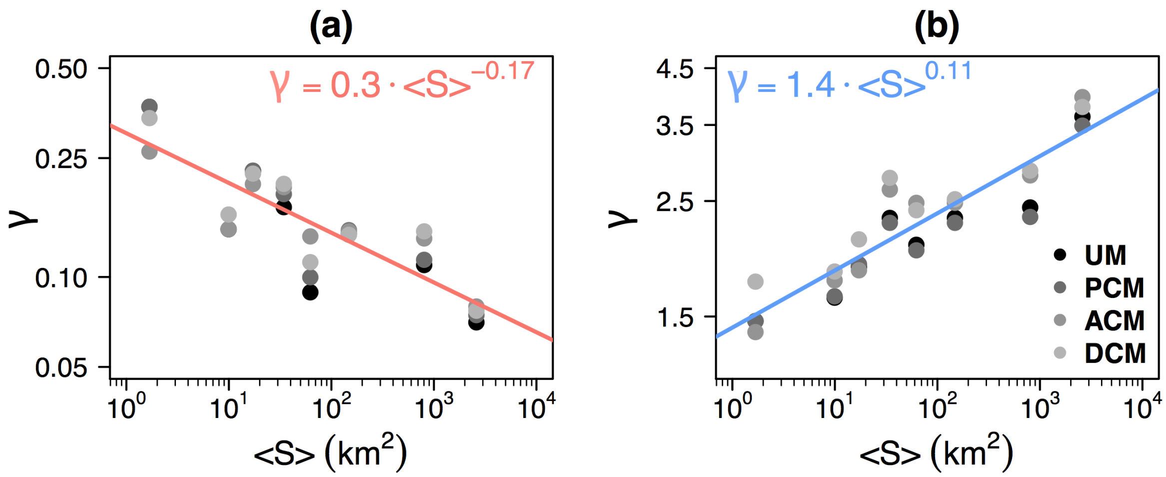


Figure 4.9: Distance parameter as a function of the average unit surface as found by Lenormand *et al.* (2016). We employ this functional relationship to estimate the mobility flows at the base granularity level. Taken from [19].

6.2 Scaling relations of the distance parameter γ

In analogy with the work of Lenormand *et al.* (2016) [19], we present here the functional relationship between the rescaled distance exponent $\tilde{\gamma}$ obtained through our optimisation method and three different distance measures: the average unit surface $\langle S \rangle$, the maximum inter-cluster d_{max} and the mean inter-cluster distance $\langle r_{\mathcal{AB}} \rangle$. Figure 4.10 shows the observed scaling. We find that our $\tilde{\gamma}$ scales linearly with both linear distance measures. An estimation of the scaling parameters with linear distance can be easily extracted from a linear regression fit and the slope α and intercept β are shown in the top panel of Figure 4.10 and summarised in Table 4.2. We also show that our rescaled parameter follows a power law when studied as a function of the surface area. By plotting the obtained values against the mean unit surface on a logarithmic scale, we recover the following functional relationship

$$\tilde{\gamma} = \alpha \langle S \rangle^{\beta}. \tag{4.4}$$

We note, however, that, although our method provides a remarkable improvement in the performance of the gravity model for the dataset examined, a systematic assessment of its validity on a more comprehensive dataset would be necessary to establish whether consistent results are obtained across different urban areas.

Measure	α	β
Average unit surface $\langle S \rangle$	1.342 ± 0.002	0.0257 ± 0.0012
Average inter-cluster distance $\langle r_{\mathcal{AB}} \rangle$	$(1.41 \pm 0.14) \times 10^{-5}$	1.00 ± 0.04
Maximum inter-cluster distance d_{max}	$(6.1 \pm 0.3) \times 10^{-5}$	1.320 ± 0.003

Table 4.2: Parameter values and their standard error extracted from the linear regression fit.

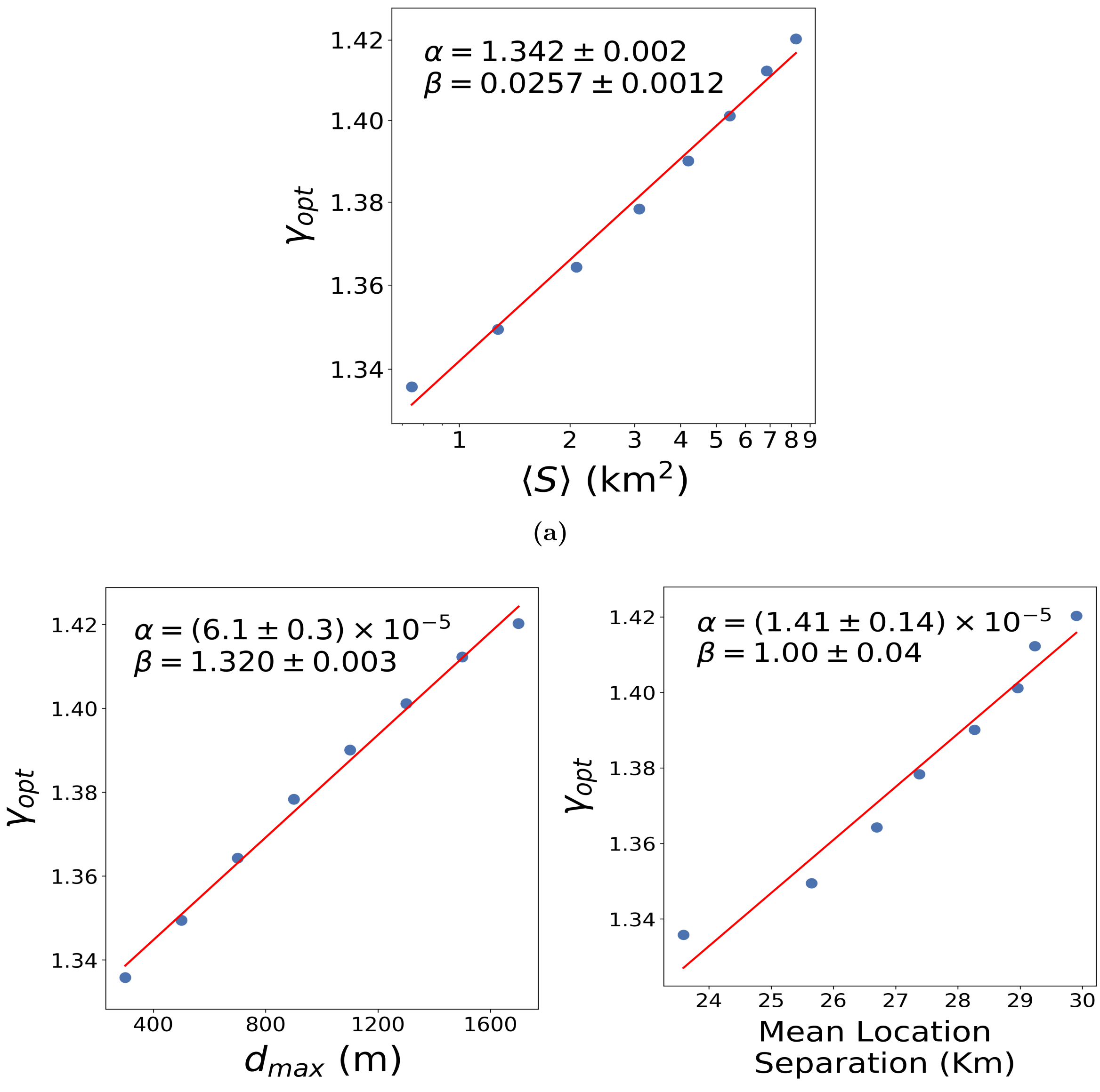


Figure 4.10: Parameter values as a function of distance and surface area. In (a), a log-log plot of $\tilde{\gamma}$ obtained by optimising the scaling error at 8 clustering levels on the London dataset shows the functional relationship with the average cluster size, which significantly differs from that found by Lenormand *et al.* (2016). (b) and (c) show the linear relationship between $\tilde{\gamma}$ and linear distance, found through the same process. In all cases the red line is the best linear fit.

Chapter 5

Conclusions

In this report we have presented a quantitative investigation of the role of spatial scales in human mobility modelling. To this end, we have adopted a dual approach: we first built the mathematical tools to compare and analyse the multiscale performance of the most widely used models, gravity and radiation, and then developed the computational simulations to reproduce this analysis and validate our theoretical predictions.

Through this analysis, we have identified a number of deficiencies in the treatment of diverse spatial scales and showed that this leads in both models to inconsistent predictions when studying the same population distribution at different resolution levels. In this respect, we have shown, in agreement with earlier studies, that the multiscale performance of the gravity model generally outperforms that of the radiation model.

The key results and contributions of our project can be summarised as follows:

- Scaling error: we proposed a new metric that encapsulates the error introduced when arbitrarily grouping spatial units and re-estimating the flows in a coarse-grained spatial system.

- Systematic comparison of the gravity and radiation model: based on our metric ϵ and on pure theoretical grounds, we provided simulations of idealised and more realistic statistical distributions of spatial units with the purpose of comparing the predicted mobility flows between different models on an equal footing.

- Optimisation procedure for the gravity model: we suggested a new approach that, given a sufficiently accurate calibration at a fine-grained population distribution, allows to tune the distance parameter without further use of empirical data, improving upon previously found empirical formulae.

Our results may therefore be useful to practitioners to gain a better understanding of the correlation between the degree of accuracy of the predicted mobility flows and the distance range considered, as well the as the granularity of the data. Our findings may also serve as a starting point for a broader investigation of the scale-dependence of other mobility models.

1. Thoughts for future research

Although successful in providing clearer insights into the scale-dependence of mobility models, our research presents some limitations: although bigger datasets comprising of global population data at a similar spatial resolution [40] were briefly explored, these were not employed for the purposes of our final analysis, due to time constraints. It would be an interesting extension of our work to employ these data sources and replicate our study on a bigger scale. Moreover, a comparison with real mobility data is necessary to complement our approach and fully validate our findings beyond the theoretical level. A more sophisticated coarse-graining procedure would also need to be devised to fully account for the high heterogeneity of different geographical distributions above the city level and in rural areas. With regards to this, a possible direction could be the implementation of an adaptive grid refinement method that selectively coarse-grains a system by correctly determining which units may be grouped together for the purposes of the flow estimation. Alternatively, coarse-graining procedures specifically designed for mobility networks, like the one proposed in [41], may be explored. Finally, a detailed comparison of the scaling error obtained by making use of different clustering techniques could be the subject of a separate study.

Another interesting direction for future research would be to exploit our approach for the optimisation of the scaling error with the aim of increasing the multiscale performance of a generic mobility framework.

All figures in the report were created by me and my project partner, unless explicitly stated.
The software package to simulate mobility flows and analyse them using the approach in this report can be found at `https://github.com/jbremz/human_mob`.

Bibliography

1. Balcan, D., Colizza, V., Gonçalves, B., Hu, H., Ramasco, J. J. & Vespignani, A. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences* **106,** 21484–21489 (2009).
2. Ravenstein, E. G. The Laws of Migration. *Journal of the Statistical Society of London* **48,** 167–235 (1885).
3. Jensen-Butler, C. Gravity Models as Planning Tools: A Review of Theoretical and Operational Problems. *Geografiska Annaler. Series B, Human Geography* **54,** 68–78 (1972).
4. Karamshuk, D., Noulas, A., Scellato, S., Nicosia, V. & Mascolo, C. *Geo-spotting: mining online location-based services for optimal retail store placement* in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (2013), 793–801.
5. Perez-Saez, J., King, A. A., Rinaldo, A., Yunus, M., Faruque, A. S. & Pascual, M. Climate-driven endemic cholera is modulated by human mobility in a megacity. *Advances in Water Resources* (2016).
6. Molodecky, N. A., Blake, I. M., O'Reilly, K. M., Wadood, M. Z., Safdar, R. M., Wesolowski, A., Buckee, C. O., Bandyopadhyay, A. S., Okayasu, H. & Grassly, N. C. Risk factors and short-term projections for serotype-1 poliomyelitis incidence in Pakistan: A spatiotemporal analysis. *PLoS medicine* **14,** e1002323 (2017).
7. Wang, Q. & Taylor, J. E. Patterns and limitations of urban human mobility resilience under the influence of multiple types of natural disaster. *PLoS one* **11,** e0147299 (2016).
8. Evans, T. S., J., R. R. & Knappett, C. Interactions in Space for Archaeological Models. *Advances in Complex Systems* **15,** 1150009 (2012).
9. González, M. C., Hidalgo, C. A. & Barabási, A.-L. Understanding individual human mobility patterns. *Nature* **453,** 779–782 (2008).
10. Song, C., Koren, T., Wang, P. & Barabási, A.-L. Modelling the scaling properties of human mobility. *Nature Physics* **6,** 818 (2010).
11. Zipf, G. K. The P 1 P 2/D hypothesis: on the intercity movement of persons. *American sociological review* **11,** 677–686 (1946).

12. Zhao, Z.-D., Huang, Z.-G., Huang, L., Liu, H. & Lai, Y.-C. Scaling and correlation of human movements in cyberspace and physical space. *Physical Review E* **90,** 050802 (2014).

13. Masucci, A. P., Serras, J., Johansson, A. & Batty, M. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E* **88,** 022812 (2013).

14. Barthélemy, M. Spatial networks. *Physics Reports* **499,** 1–101 (2011).

15. Yan, X.-Y., Wang, W.-X., Gao, Z.-Y. & Lai, Y.-C. Universal model of individual and population mobility on diverse spatial scales. *Nature Communications* **8** (2017).

16. Yang, Y., Herrera, C., Eagle, N. & González, M. C. Limits of predictability in commuting flows in the absence of data for calibration. *Scientific Reports* **4** (2014).

17. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484,** 96–100 (2012).

18. Stefanouli, M. & Polyzos, S. Gravity vs radiation model: two approaches on commuting in Greece. *Transportation Research Procedia* **24.** 3rd Conference on Sustainable Urban Mobility, 3rd CSUM 2016, 26 – 27 May 2016, Volos, Greece, 65–72 (2017).

19. Lenormand, M., Bassolas, A. & Ramasco, J. Systematic comparison of trip distribution laws and models. *Journal of Transport Geography* **51,** 158–169 (2016).

20. Chen, Y. The distance-decay function of geographical gravity model: Power law or exponential law? *Chaos, Solitons & Fractals* **77,** 174–189 (2015).

21. Erlander, S. & Stewart, N. F. *The gravity model in transportation analysis: theory and extensions* (Vsp, 1990).

22. Flowerdew, R. & Aitkin, M. A method of fitting the gravity model based on the poisson distribution. *Journal of Regional Science* **22,** 191–202 (1982).

23. Bhattacharyya, P. & Chakrabarti, B. K. The mean distance to the nth neighbour in a uniform distribution of random points: an application of probability theory. *European Journal of Physics* **29,** 639 (2008).

24. Yan, X.-Y., Zhao, C., Fan, Y., Di, Z. & Wang, W.-X. Universal predictability of mobility patterns in cities. *Journal of The Royal Society Interface* **11,** 20140834 (2014).

25. Liang, X., Zhao, J., Dong, L. & Xu, K. Unraveling the origin of exponential law in intra-urban human mobility. *Scientific reports* **3,** 2983 (2013).

26. Singleton, A. *2011 Population Weighted Centroids (LSOA/Data Zone)* `https://data.cdrc.ac.uk/dataset/cdrc-2011-population-weighted-centroids-gb`.

27. *Census geography* Office for National Statistics. `https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography`.

28. Thiemann, C., Theis, F., Grady, D., Brune, R. & Brockmann, D. The structure of borders in a small world. *PloS one* **5,** e15422 (2010).

29. Grauwin, S., Szell, M., Sobolevsky, S., Hövel, P., Simini, F., Vanhoof, M., Smoreda, Z., Barabási, A.-L. & Ratti, C. Identifying and modeling the structural discontinuities of human interactions. *Scientific reports* **7,** 46677 (2017).

30. Grafarend, E. *The optimal universal transverse Mercator projection* in *Geodetic Theory Today* (1995), 51–51.

31. *Lower Layer Super Output Areas (December 2011) Generalised Clipped Boundaries in England and Wales* `http://geoportal.statistics.gov.uk/datasets/da831f80764346889837c72508f046fa_2`.

32. Müllner, D. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378* (2011).

33. Eads, D. *Hierarchical clustering (scipy.cluster.hierarchy)* The Scipy community. `https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html`.

34. Wong, D. W. S. in *WorldMinds: Geographical Perspectives on 100 Problems: Commemorating the 100th Anniversary of the Association of American Geographers 1904–2004* (eds Janelle, D. G., Warf, B. & Hansen, K.) 571–575 (Springer Netherlands, Dordrecht, 2004).

35. Eads, D. *Hierarchical clustering (scipy.cluster.hierarchy.linkage)* The Scipy community. `https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.linkage.html`.

36. Wesolowski, A., O'Meara, W. P., Eagle, N., Tatem, A. J. & Buckee, C. O. Evaluating spatial interaction models for regional mobility in sub-Saharan Africa. *PLoS computational biology* **11,** e1004267 (2015).

37. Kang, C., Liu, Y., Guo, D. & Qin, K. A generalized radiation model for human mobility: spatial scale, searching direction and trip constraint. *PloS one* **10,** e0143500 (2015).

38. Chen, Y. & Huang, L. A scaling approach to evaluating the distance exponent of the urban gravity model. *Chaos, Solitons Fractals* **109,** 303–313 (2018).

39. Nocedal, J. & Wright, S. J. Quasi-newton methods. *Numerical optimization,* 135–163 (2006).

40. *High Resolution Settlement Layer (HRSL)* Source imagery for HRSL © 2016 DigitalGlobe. Facebook Connectivity Lab and Center for International Earth Science Information Network - CIESIN - Columbia University. `https://ciesin.columbia.edu/data/hrsl/`.

41. Louail, T., Lenormand, M., Picornell, M., Garcıa Cantú, O., Herranz, R., Frias-Martinez, E., Ramasco, J. J. & Barthelemy, M. Uncovering the spatial structure of mobility networks. *Nature Communications* **6** (2015).

Appendix A

The normalisation factor in the gravity model

1. A numerical validation

We carried out numerical tests to establish the validity of the approximations made in the derivation of ϵ in the gravity model (Eq. (2.10)). Among these, particularly important is they key assumption that the normalisation constants in the pre- (k_i) and post-clustering ($\tilde{k}_i$) phase can be considered to be approximately equal. As shown in Fig. A.1, we find that a distribution of $N = 100$ locations, as used in our simulation, is sufficient to allow for this approximation without introducing significant bias and offers a fair representation of the asymptotic limit ($N \gg 1$).

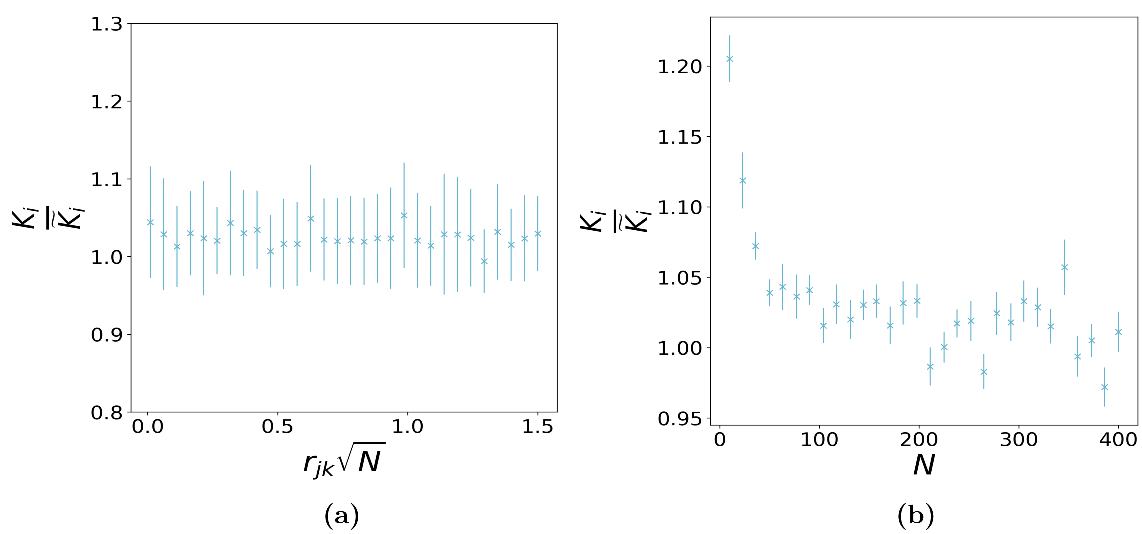
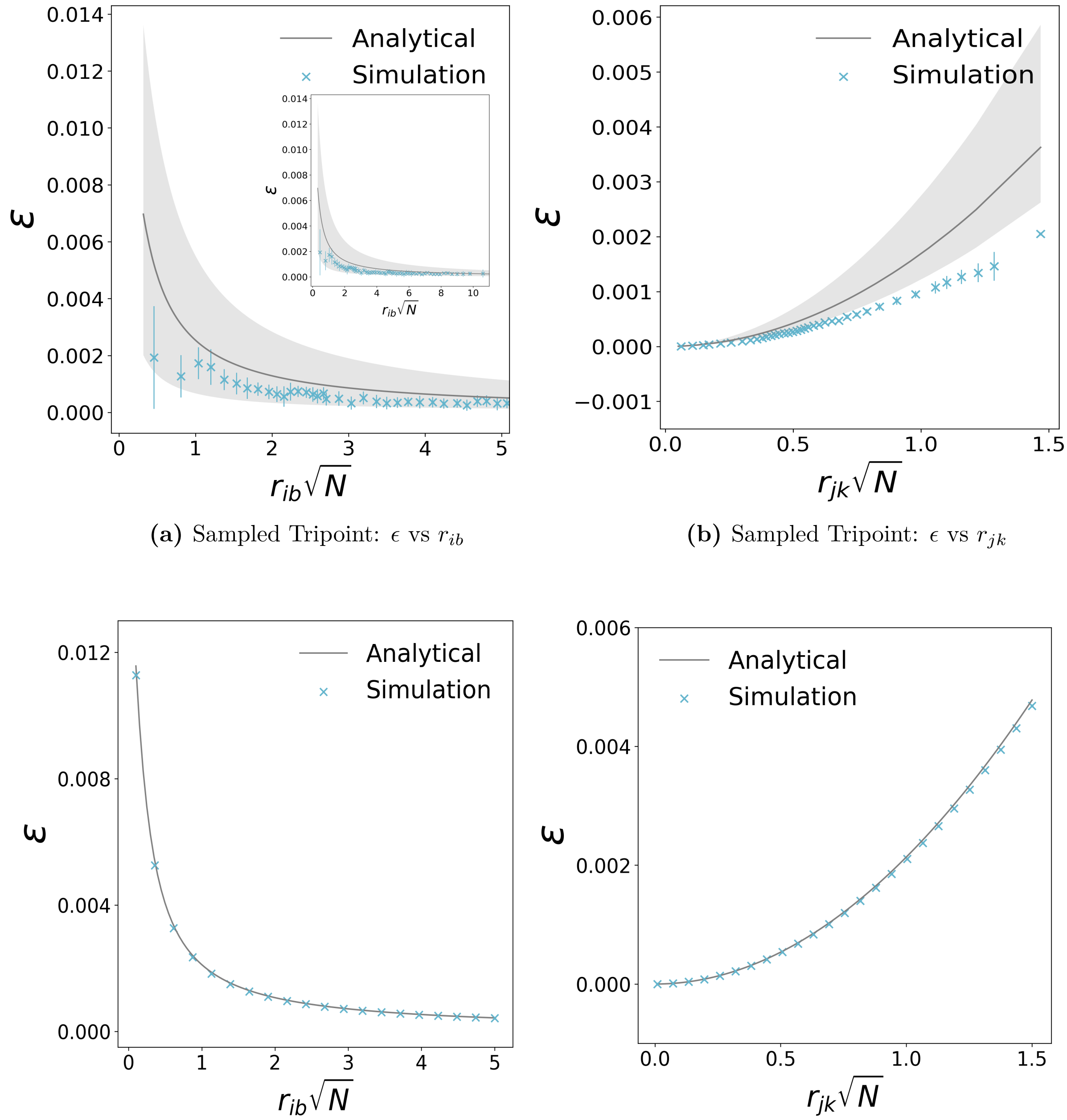


Figure A.1: Ratio of the normalisation factor in the exponential gravity model at the pre-clustering (k_i) and post-clustering ($\tilde{k}_i$). In (a) the ratio is plotted as a function of the intra-cluster distance r_{jk}, showing a negligible difference between the normalisation factors across the whole range. In (b) the ratio is plotted as a function the total number of units N, indicating that the factors are approximately equal provided that $N \gtrsim 100$.

Appendix B

Tripoint scaling error in the exponential gravity model

Similarly to to Chapter 3 2.1, we present here the results for the scaling error ϵ obtained in the tripoint configuration by using an exponential distance decay function in the gravity model.



(c) Explicit Tripoint: ϵ vs r_{ib}

(d) Explicit Tripoint: ϵ vs r_{jk}

Figure B.1: Relationship of the scaling error with distance in the exponential gravity model. The blue markers indicate the numerical results obtained through the sampled tripoint algorithm in (a)-(b) and the explicit tripoint algorithm in (c)-(d), plotted as a function of r_{ib} (left) and r_{jk} (right). The grey curve represent the analytical error ϵ. According to the functional relationship in [19], we use $\gamma = 0.69$. All other parameters are the same as in Ch. 3 Sec. 2.1. The scaling error is here below 0.02% even at considerably small origin-destination separation, when this is comparable to the typical inter-site distance ($r_{ib} \simeq 1$). Therefore the exponential gravity model outperforms the power-law form by yielding an error that is 2 orders of magnitude smaller in the range considered.

Appendix C

The role of population density

In order to fully validate our tripoint method, in which we treated the population distribution as uniform in space, we investigated what effect a non-uniform distribution has on the scaling error associated with the coarse-graining hierarchy described in Chapter 4.

One could in fact argue that the heterogeneity of the population distribution might introduce an additional variable compared to the simplified model of the tripoint system, thus rendering the two fundamentally different. We first note that, in virtue of the fact that, by definition, ϵ is a fractional error, we expect the population to play a marginal role in determining the magnitude of the error. Figure C.1 provides evidence that is is indeed the case and therefore further corroborates our approach.

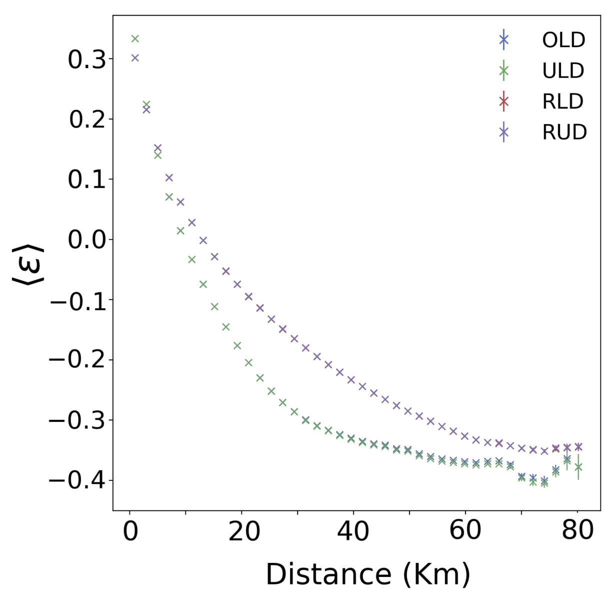


Figure C.1: Average scaling error at a coarse-graining level corresponding to a maximum inter-cluster distance $d_{max} = 1000m$ (London dataset). The error is computed using the original London distribution (OLD) of locations, the same spatial distribution but with uniform masses (ULD), the mass distribution randomly assigned to the London location distributions LD (RLD), and then a random distribution with uniform masses (RUD). While the spatial location distribution affects the scaling error, the difference between a heterogeneous and a homogeneous population distribution is negligible.