Network Models of Spatial Interactions, Human Mobility and Navigation Ability

Rodrigo Leal Cervantes

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2022

Η Ιθάκη σ΄ έδωσε το ωραίο ταξίδι.

Κ. Π. Καβάφης

Ithaca gave you the beautiful journey.

Coming to the end of my doctorate, I find even more meaning in these words. Those who travelled with me already know that I found the journey to be long and challenging, but it is also true that Ithaca gave me something beautiful.

I want to thank my supervisors, Renaud and Sam, who guided and motivated me. From the scholars, I learned and learned. I will treasure my visits to Sam's office with the most amazing collection of mathematical objects, and the dinners and visits to the pub with Renaud, who was always warm and supportive.

I thank my examiners, Elsa and Xiaowen, who made me look at my work with fresh eyes. I really enjoyed my Viva because it felt like a conversation with colleagues who cared about providing ideas to improve the work I was presenting to them.

The most beautiful riches that I gained were my fellow travellers: Blas, Leo, Carlota and Santiago who became my family; my MMSC friends, particularly Jack, Ilan, Reka, Steffen, John, Thomas, and Giancarlo and Giulia with whom I also shared travels; my academic siblings Karel, John, and Yu; and Samu and Lionel who visited the group and who shared so much with us.

I want to thank the whole InFoMM family, students, academics and administrators. I will have very fond memories of coffee at 11 with them. With my cohort I lived through so many things, both good and bad, and I am grateful that I had them along the way. I want to thank Amanda who was so kind, and Hattie who was always so caring and helpful to my academic siblings and me.

Alonso, Tere, Ana and Carmen have travelled with me since before I had even started this journey, and I owe a great deal to them. They put Ithaca on my mind.

And amongst everyone else, I want to thank Natalia who has been journeying with me more than six years and who was my greatest support when I did encounter Cyclops and Laistrygonians. I have sailed into many ports with her, and my wish for us is that Ithaca, this Ithaca, has only been one more stop and there will be other Ithacas.

Abstract

Networks provide a useful mathematical model for systems that are composed of agents, also called nodes, that interact in a pairwise fashion. Inside a spatial network, nodes are embedded in space and they typically interact much more strongly within their close surrounding, resulting in a special pattern of connections. Geographers, physicists and applied mathematicians have a set of tools at their disposal to understand these systems, and there is a close relationship between spatial networks and population-level models that were originally meant to represent human mobility, but that have been generalised to model more abstract spatial flows.

Over the last decade, there has been an effort to adapt network tools for community detection — whose aim is to describe a system in terms of its mesoscale organisation into groups or communities — to spatial networks. So far, the methods that have been proposed are based on the modularity function, a heuristic function that compares network partitions against a suitable null model. However, a mesoscale description that is gaining traction relies instead on the stochastic block model (SBM), a generative model that can be fitted to an observation using statistical inference. We propose a methodology to leverage an SBM for the case of spatial networks.

We are guided in this by an application to study a network linking customers to the stores they have shopped in, built from anonymised shopping records belonging to a large UK retailer. We also characterise customer shopping behaviours, specially from the lens of their mobility. Lastly, we study a dataset that was collected to assess spatial navigation, and we propose new metrics that are used in comparing the trajectories of healthy individuals, at-genetic-risk patients and patients diagnosed with dementia. Our hope is that this early proposal can be refined in the future by medical practitioners and used to detect early-onset Alzheimer's Disease.

Contents

Chapter 1

Introduction

Many aspects of our connected lives are shaped by space. This fact is obvious to anyone who has boarded a plane to visit a friend on the other side of the ocean only once or twice in their lives, but regularly sees other friends in the same city. Long-distance friendships exist thanks to the wonders of air-travel and telecommunications, but there is an opportunity cost to pay when we are further away from our friends.

This thesis has as its main objects of study different examples of spatial networks as well as other systems for which the modelling ideas of spatial networks can be applied. A network is a mathematical model of a system composed of agents that interact in a pairwise fashion. By focusing on the patterns formed by the connections — the topology of the connections — non-spatial networks have provided us with many great insights of various complex systems. For a system that is embedded in space, however, a description that is purely topological might miss crucial aspects of how the system behaves because space will have an influence, either directly or indirectly through the higher cost of distant interactions [11].

We have encountered a spatial network in the example of the friendship network with its opportunity cost. Infrastructure, be it transportation (air-traffic, road or rail) or telecommunication networks (the servers and switches that make up the Internet) are further examples and they are shaped by a monetary cost. A vascular system [83] or the growth of slime mould [118] are also modelled as spatial networks, taking their form from an energetic cost. The structure of all of these systems is affected by spatial constraints and it will be important to know the location of the agents in order to model the interaction costs.

Thinking about the tools that have been developed over more than two decades in Network Science, there is work to be done when it comes to spatial networks so that the tools can take advantage of information other than the purely topological. Paradoxically, it was by abstracting space away that Euler solved the problem of the seven bridges of Königsberg and founded the discipline of Graph Theory, from which our field developed. Yet, more than two centuries later, we argue for a re-introduction of space into a larger share of network models.

Some of the systems that we concern ourselves with in the present work are a collection of shopping records from a large retailer, commuting at a national level in the United Kingdom, and navigation inside a virtual environment. Methodologically, we propose a way to describe

the mesoscale structure of a spatial network, we seek to cluster similar patterns of human mobility and shopping habits, and we develop performance metrics to determine an individual's navigation ability.

1.1 Spatial systems studied in this thesis

As a thesis in Applied Mathematics, this work places a great deal of emphasis on the applications themselves. As such, many of the research questions that will be explored are linked to specific datasets or data sources that we now introduce. What links all of them is the spatial nature that they have, and thus that they are amenable to be studied using the framework of a spatial network.

1.1.1 Records of retail baskets

Most people have a preferred retailer to shop for groceries which they normally visit across multiple physical stores, and possibly also online. Often, the stores have varied characteristics (such as parking space), a different format (convenience, hypermarket, etc.) and a different assortment of products. The result is that each store fulfils particular shopping needs and customers shop across a *network* of stores belonging to their preferred retailer.

Through our industrial partner, Dunnhumby, we were given access to anonymised information about baskets shopped by loyalty card holders of a large retailer in the United Kingdom (henceforth referred to as the *Retailer*).

The massive amounts of anonymised data that are made available through loyalty programmes have revolutionised the industry by shining a light on many aspects of consumer habits and informing corporate decisions about product assortment and personalised promotions and recommendations. Nevertheless, analysing such a rich dataset — with its temporal dimension, its spatial dimension and the almost infinite subtleties available when we look at the products being bought — is a complex task.

We are interested in the spatio-temporal dimensions of the dataset and we model the data points using a network connecting customers to the stores they visit: the *customer-store network*. Space reveals itself through the network structure: clearly, the connections must be shaped by the customers' proximity to the stores. One of our contributions will be a methodology to study the network after we have "discounted" the spatial effects. We also compute time series and we analyse the seasonal differences in shopping habits — including during the period of lockdown restrictions due to the COVID-19 pandemic.

1.1.2 Commuting in the United Kingdom

We complement the spatial analysis of the retail dataset by also looking at the network of commuting in the United Kingdom as recorded in the 2011 census.[1] This dataset covers the same geographical region as the retail network and as such it serves as a useful point of comparison.

[1] At the moment of writing, the first results of the 2021 census have been published but commuting is not included yet and the results will be available in 2023.

In the census, respondents were asked to name their place of residence and work. We use the published results that have been aggregated at the level of 404 local authorities, the administrative divisions of the United Kingdom which correspond to councils, metropolitan districts and boroughs (called different names in England, Scotland, Wales and Northern Ireland).[2]

1.1.3 Navigation trajectories in a virtual environment

In the last few years, medical researchers studying dementia at the University of East Anglia and University College London have been developing a virtual game with the purpose of helping them quantify navigation ability in a clinical setting. The game — called Sea Hero Quest — can be downloaded to mobile devices and it has been played by millions of volunteers around the world. The data collected takes the form of trajectories in constrained settings.

Our colleagues shared with us the subset of trajectories recorded for volunteers in the United Kingdom, as well as basic (self-reported) demographic information such as gender and age. We also had access to the data generated by clinical patients that were part of an at-genetic-risk study for Alzheimer's disease and we study both data sources in relation to each other.

Having introduced the spatial systems that are studied throughout the thesis, we now set forth the research questions that they motivate and will make up the bulk of our work.

1.2 Revisiting human mobility models

Researchers have studied human mobility for well over a hundred years [97]. The pursuit of a universal law that governs how people move has fascinated social scientists, physicists and applied mathematicians alike [39, 99, 111, 116, 117, 130]. At the same time, characterising the movements of humans has important applications for forecasting and transport planning and in the last decade the wealth of geo-localised information collected has made the topic more prominent [9].

Models of human mobility aim to capture various aspects of real human beings commuting to work, going shopping, travelling, and more generally going about their daily lives. One key goal is to capture regularities both in time and space. While it might be challenging to predict the movement of individuals, there are typically predictable patterns in the collective behaviour of many thousands of human beings.

In Chapter 3, we re-introduce and extend a unified framework that builds upon aggregate human mobility models. Our goal is to reinterpret them as statistical models, for one because we recognise the stochastic nature of the systems being modelled, but most specially because (in Chapter 4) we build upon significance tests to analyse spatial networks. To the best of our knowledge, our work provides the first use case of the doubly-constrained model (which will be introduced at the beginning of Chapter 2) as a statistical model.

[2]For boundaries and details about local authorities by the Office for National Statistics, see https://www.statistics.digitalresources.jisc.ac.uk/dataset/2011-census-geography-boundaries-local-authorities.

1.3 The mesoscale organisation of spatial networks

Community detection (CD), the problem of finding the mesoscale organisation of a network [95], has emerged as one of the most important tools in Network Science. Each discipline within the field, be it Applied Mathematics, Statistical Physics, Computer Science or Sociology, has contributed their own methods. We review the literature of community detection in Chapter 2.

The characterisation of the organisation at an intermediate level — a level larger than the individual agents but smaller than the system as a whole — has important applications like description, summarisation, visualisation, and partitioning of a network. According to Newman [77], community detection provides a "deeper insight into the *function* of networked systems".

Here, we focus on the community detection problem for the customer-store and the commuting network, two examples of spatial networks in which distant locations are less likely to interact, resulting in a special network topology. Most community detection methods have an underlying model of structure for the network studied, so the models need to be adapted to incorporate spatial information.

So far, the methods that have been proposed are based on modularity maximisation which identifies communities as groups of nodes that are more strongly connected than expected, at least compared to a (spatial) null model. These methods, however, inherit the limitations of the modularity function like being restricted to find assortative mesoscale structure and more importantly, overfitting the data.

We propose instead a methodology to adapt the stochastic block model (SBM) — a family of parametric models to which statistical inference is done — to the spatial case. This offers several advantages, like being able to easily incorporate spatial models that are directed, allowing for more varied types of mesoscale structure, and giving stronger guarantees that in the absence of statistical evidence the method will not partition the network into an arbitrarily large number of groups.

In Chapter 3, we lay the grounds and we introduce a pair of human mobility models that aim to capture the regularity in the spatial networks. Then, in Chapter 4, we employ the statistical interpretation of the aforementioned mobility models to extract a signal in the form of a network that we call the spatial backbone to which we can fit a layered version of the SBM. With the stochastic block model, we obtain a grouping of the nodes and a mesoscale description that has been "corrected" for spatial constraints to reveal other factors that contribute to the underlying structure of a spatial network. In doing so, we find cultural and economical signals shaping the commuting network and an organisation of the retail network based on the type and size of the different stores.

1.4 Spatio-temporal analysis of shopping habits

The space-corrected SBM will be applied successfully to study the interactions between different stores. Nevertheless, our methodology — and the fact that we are unable to give coordinates to households — forces us to aggregate the visits of individual customers and to study only the

stores (in technical terms, we project the customer-store bipartite network to keep only the store nodes). Being unhappy with having such an incomplete picture, we also turn our attention to the trail left behind by the customers when they visit different stores.

There is a good amount of information that can be inferred about the habits of customers from a high-level analysis of the anonymised shopping records. Information such as the spending, the number of items bought, the distribution of the shopping during the days of the week and across the different stores, all of this can be computed cheaply and efficiently at a basket level, without the need to look inside the baskets. At the same time, this data already carries plenty of information about seasonal or regional patterns and studying them will be the subject of the first few sections inside Chapter 5.

Furthermore, the pandemic of COVID-19 created a shock that we can analyse to quantify how shopping for groceries — one of the most important activities we were able to do during lockdown — was affected. We know that the visits to most supermarkets decreased (we probably have first-hand knowledge of this since many of us tried to visit them less frequently) but where did it decrease most significantly? Was there a demand transfer and were there stores that increased their sales?

We answer positively to this question, and this then leads us to the interesting problem of reconstructing how demand could have flowed, once the lockdown was put in place, from stores with decreased demand to the stores with increased demand. We pose and then solve a transport problem which can be efficiently calculated at the level of individual customers. When aggregated across different locations, we get a reconstruction of the demand transfer that is applicable to the COVID-19 shock to the retail industry, but that could easily be generalised to other settings.

Lastly, our final contribution in Chapter 5 is exploring whether the shopping records can be a useful data source to study human mobility or whether the resolution they provide might be too coarse compared to GPS or call detail records.

1.5 Spatial navigation for early detection of Alzheimer's disease

Turning our attention to a different problem, we look in Chapter 6 at the application of spatial networks and particularly a representation called the *origin-destination* matrix to determine erratic navigation in a virtual game.

Recent studies have shown that problems with navigation could serve as an early indicator of pre-clinical Alzheimer's disease. However, determining what is to be considered abnormal navigation needs careful consideration because of the natural inter-individual variability due to many factors other than a cognitive impairment.

For this reason, medical researchers developed a virtual game that can be downloaded to mobile devices and played by volunteers around the world. The data collected serves as a large normative dataset, and we use it to quantify the navigation ability of an individual compared to a large cohort with matching demographic characteristics. We propose new metrics that our medical colleagues have not studied before, and that have the potential to distinguish more

clearly problematic navigation ability.

1.6 Statement of originality

The work contained in this thesis is original and was carried out by myself except where explicitly stated. My two supervisors provided me with ideas and guided me in my research, but I produced the derivations and the methods presented here.

Chapter 6 is special in that it was a collaboration with another DPhil student and a group of medical colleagues. The medics provided the data and told us it would be a good idea to compare different cohorts of clinical patients, but the fundamental contribution of developing new geometric metrics to quantify the navigation ability is original. My colleague proposed the curvature and boundary affinity metrics while all of the metrics derived from the OD matrix are my own. The ideas of employing the visiting order and the way to study the discerning power of the metrics comparing the percentiles of the genetic groups are also mine. My colleague performed some of the analyses herself and whenever her results are included because they would add to the discussion (mainly as figures) this is indicated.

There are two of articles being prepared from the work presented in Chapters 4 and 6:

- Rodrigo Leal-Cervantes, Sam Howison and Renaud Lambiotte, *Stochastic Blockmodelling of Spatial Networks*

- Rodrigo Leal-Cervantes, Uzu Lim, Gillian Coughlan, Heather Harrington, Renaud Lambiotte, Hugo Spiers and Micheal Hornberger, *Geometric Measures of Navigation in 61,000 Adults and Their Application to Alzheimer's Disease*

We estimate the progress for each to be around 60% and 90% respectively.

Chapter 2

Preliminaries

In this chapter, we introduce the notation of networks and we discuss the relevant work that has been done in the field of community detection. We also introduce very briefly some aspects of the bipartite customer-store networks that is then projected (for reasons that are explained later in the thesis) and studied further in Chapters 3 and 4.

2.1 Basic definitions

A *network* $G = (V, E)$, also called a graph, is a mathematical object that represents the pairwise patterns of relationships between entities [77]. The various types of entities — persons, proteins, power stations, etc. — are abstracted and represented by the set of *nodes*, V, and the relationships are encoded inside the set of *edges*, $E \subseteq V \times V$. We denote by $N = |V|$ the number of nodes and the number of edges is $M = |E|$.

In a *directed* network, an edge $e \in E$ is an ordered pair (i, j) that indicates that node $i \in V$, the *source*, is connected to node $j \in V$, the *target*. We say that nodes i and j are *neighbours* and furthermore, relative to a selected node i, we distinguish the *in*-neighbours $\{s \in V : (s, i) \in E\}$ and the *out*-neighbours $\{s \in V : (i, s) \in E\}$. An edge (i, i) that connects a node to itself is called a *self-edge*.

A network is called *undirected* if for every edge $(i, j) \in E$ the reciprocal (j, i) is also in E. Relative to node i, the neighbours are denoted $N(i) = \{s \in V : (i, s) \in E\}$. When the context is clear, we use the notation $j \sim i$ as a shorthand for $j \in N(i)$.

The topological density is defined as the fraction of realised edges over the number that is possible, $\rho_t = M/N(N-1)$ for directed networks (excluding self-edges) or $\rho_t = 2M/N(N-1)$ for undirected networks. By definition, this measure satisfies $\rho_t \in [0, 1]$. In both of the previous expressions we exclude self-edges because this is in line with the network models used in Chapter 4, but the denominator can obviously be changed according to the network model being employed and, for example, we would use N^2 for a directed network where self-edges are present.

In some applications it is useful to quantify the strength of the relationships between nodes. This is achieved by associating a *weight* w_{ij} to the edge (i, j) (we assume that an undirected network has $w_{ij} = w_{ji}\ \forall\, i, j$) and the resulting network is called *weighted*. An *unweighted*

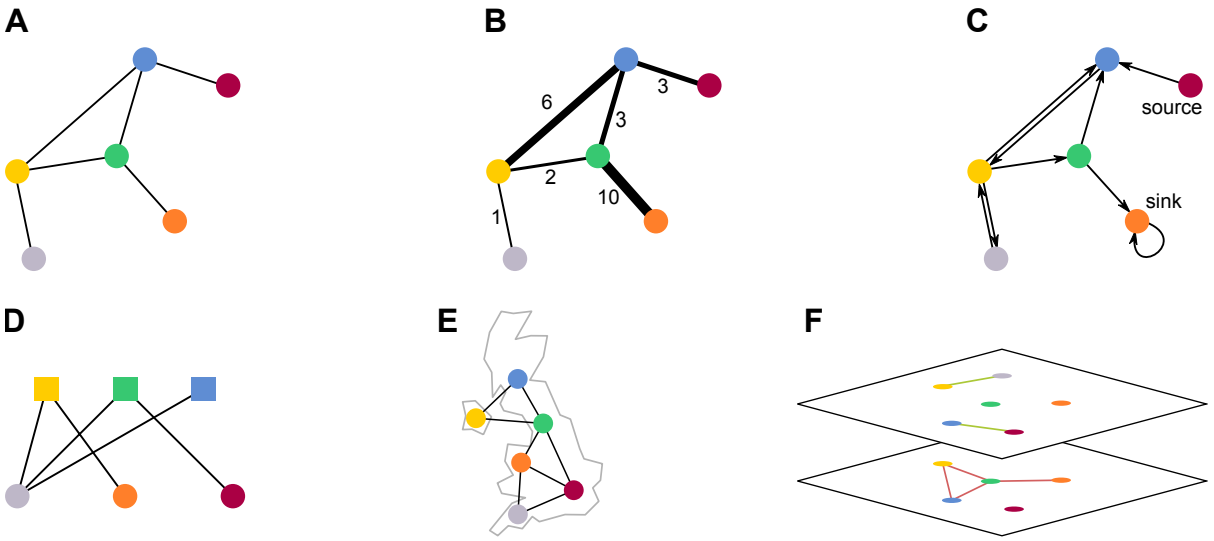


Figure 2.1: Different types of networks: **A.** undirected and unweighted, **B.** weighted, **C.** directed, **D.** bipartite, **E.** spatial and **F.** layered with $C = 2$ layers.

network can be thought in terms of the special case $w_{ij} = 1 \, \forall (i,j) \in E$.

If we select an ordering for the set of nodes V, a useful way of representing a network is by means of the *adjacency matrix*, $\mathbf{A} = (A_{ij}) \in \mathbb{R}^{N \times N}$, defined as

$$A_{ij} = \begin{cases} w_{ij} & \text{if } (i,j) \in E, i \neq j \,, \\ q \, w_{ii} & \text{if } (i,i) \in E \,, \\ 0 & \text{otherwise} \,, \end{cases} \tag{2.1}$$

where the factor in front of w_{ii} is $q = 1$ for directed networks, and is set to $q = 2$ by convention for undirected networks (so that the calculations involving sums of edges can be consistently extended to include self-edges).

In an undirected and unweighted network, the *degree* k_i of node i counts the number of edges adjacent to i, or equivalently, the number of its neighbours, $|N(i)|$. This can be easily generalised to the weighted case using the row (or column) sums of the adjacency matrix, and we typically call this the *strength* of node i,

$$k_i = \sum_{j \sim i} A_{ij} \,. \tag{2.2}$$

In matrix form, the vector of degrees (strengths) is given by $\mathbf{k} = A\mathbb{1}$, where $\mathbb{1} \in \mathbb{R}^N$ is the vector of ones.

For directed networks we need to distinguish between the in-degree and out-degree (in-strength and out-strength), calculated respectively as $\mathbf{k}^{\text{in}} = A^\top \mathbb{1}$ and $\mathbf{k}^{\text{out}} = A\mathbb{1}$. A node with no incoming edges $k_i^{\text{in}} = 0$ is called a *source* and one with no outgoing edges $k_i^{\text{out}} = 0$ is called a *sink*.

Figure 2.1 shows a schematic representation for the different types of networks that we presented so far and others that will be introduced shortly.

On a network, a distribution $p : V \to [0, 1]$ is a function that assigns a non-negative number $p(i)$ to each node, satisfying $\sum_{i \in V} p(i) = 1$.

2.1.1 Varieties of network models

We now introduce other types of networks that build upon the basic building blocks that we have seen so far, directed, undirected, weighted, and that are used in different parts of the thesis.

Bipartite networks

A network is called *bipartite* if the set of nodes V can be decomposed into two disjoint sets X and Y so that edges only connect nodes in different sets: $\exists\, X, Y :\ V = X \cup Y,\ X \cap Y = \emptyset$ and for all $e = (x, y) \in E :\ x \in X,\ y \in Y$. An example can be seen in Figure 2.1D.

Typically, the use of bipartite networks makes the most sense with an undirected network, and in this case the adjacency matrix takes the special form

$$\mathbf{A} = \begin{pmatrix} \mathbf{0} & \mathbf{B} \\ \mathbf{B}^\top & \mathbf{0} \end{pmatrix} . \tag{2.3}$$

The matrix $\mathbf{B}$ of size $|X| \times |Y|$ is called the biadjacency matrix, and it fully encodes the undirected pattern of connections. Often, this smaller matrix is used instead of $\mathbf{A}$.

A bipartite network is sometimes simplified by doing a *projection* to one of the two sets of nodes. Without loss of generality, suppose that we are interested in the set X of nodes. Our objective is to construct a new network that contains only the nodes in X and where the nodes are connected if they share one or more common neighbours in Y.

Normally, we give weights to the new edges in order to preserve as much information as possible from the original bipartite network. A standard choice is the number of common neighbours shared by nodes: the X-projection is obtained using $\mathbf{B}\mathbf{B}^\top$ and the Y-projection is $\mathbf{B}^\top\mathbf{B}$. Note that the new networks are undirected. An alternative projection that produces a directed network is the method of resource allocation, proposed by Zhou et al.[129].

Multigraphs

A *multigraph* is a network that has edges with multiplicity, also called *multi-edges*. A multi-edge can be interpreted as a repeated edge between the same pair of nodes — this is why they are sometimes called parallel edges — but it can also be related to weighted networks that have integer weights. Indeed, we can think of an entry in the adjacency matrix $A_{ij} = 2$ either as a double edge or as a single edge with a weight $w_{ij} = 2$. We make use of this duality to generate synthetic networks in Chapter 4.

Spatial networks

A *spatial network* is a network that is embedded in space [10], normally two-dimensional Euclidean space (see Figure 2.1E), but also higher-dimensional or hyperbolic space. The pattern of connections, encoded in the adjacency matrix, is not enough to fully characterise the network

and we also need to know the position of all the nodes. A spatial network is called *planar* if it can be represented in two dimensions in such a way that its edges do not intersect. Spatial constraints normally lead to a distinctive network topology, and specially for planar graphs, this means a low number of high-degree nodes and a degree distribution that is not fat-tailed [11].

Layered networks

There are situations in which a system is too complex for a single network to capture the whole range of interactions present and instead a "network of networks" would be desirable. For example, temporal networks can be studied as a collection of snapshots, each capturing the interactions that took place at specific times or during specified time-windows. Social scientists sometimes view social groups from the lens of multiple interactions — familiar, amicable, professional, etc. In both cases, *multilayer* networks provide a "network of networks" modelling framework [53] that is very rich.

A full introduction to this large topic is outside the scope of our discussion, but we simply highlight one type of multilayer network: a *multiplex* network is a node-aligned collection of C edge-coloured networks represented by the set of matrices $\{A_{ij}^l\}_{l=1,\dots,C}$. The fact that nodes are aligned means that the set of nodes V is the same for all of them, and they are globally indexed across layers; the colouring of the edges means that we can tell them apart if they are present in different layers. An example with $C = 2$ layers is represented in Figures 2.1D.

2.1.2 Network partitions

A partition of the network into B non-overlapping groups, typically called communities or blocks, is represented by the vector $\mathbf{b} \in \mathbb{N}^N$ where each element satisfies $b_i \in \{1, \dots, B\}$. We let $\mathbf{e} = (e_{rs})$ denote the $B \times B$ matrix of edge counts between groups, with each element defined as

$$e_{rs} = \sum_{i,j} A_{ij}\, \delta_{b_i r}\, \delta_{b_j s}\,, \tag{2.4}$$

where the Kronecker delta, δ_{ij}, is 1 if $i = j$ and zero otherwise. We might also want to do a soft partitioning of the network in which the groups of nodes can be overlapping but there are no examples of this in the thesis, and we do not discuss the soft partitioning further.

The matrix $\mathbf{e}$ can be viewed as the adjacency matrix of a multigraph composed of B nodes, one node for each community, and the original number of edges M. Consistently with our convention for self-edges, e_{rr} stores twice the number of edges with both ends inside community r if the original network is undirected.

We define as well

$$e_r = \sum_s e_{rs} = \sum_i k_i\, \delta_{b_i r}\,, \tag{2.5}$$

the number of endpoints of an edge (also called half-edges or stubs) inside community r. This quantity is sometimes called the *volume* of r [32]. For the multigraph constructed using matrix $\mathbf{e}$ with nodes corresponding to communities, e_r can be viewed as the degree of node r.

Similarity measures

In order to assess how a particular community detection method performs, we need tools to measure the similarity (or dissimilarity) between partitions.

For two partitions $\mathbf{b}_X$ and $\mathbf{b}_Y$ of the same network, let N_r^X denote the number of nodes in community r according to partition $\mathbf{b}_X$ (and equivalently for partition $\mathbf{b_Y}$), and let N_{rs} count the number of nodes in community r of $\mathbf{b}_X$ that are also found in community s of $\mathbf{b}_Y$. The contingency table is denoted $\mathbf{N} = (N_{rs})$ and, in the context of comparing partitions, is sometimes also called the confusion matrix.

When the partitions only contain two groups indexed 0 and 1, a simple measure of agreement is given by the Jaccard index, $J(\mathbf{b}_X, \mathbf{b}_Y) = N_{11}/(N_{01} + N_{10} + N_{11})$. However, this definition has an important drawback for the application of comparing network partitions because it assigns a different meaning to the values 0 and 1 (typically, the Jaccard index is used to measure classification accuracy and the values indicate failure and success respectively).

To us, a partition represents a colouring of the nodes. We encode it with integer values for convenience but the label that a particular group is assigned should not matter; as such, we expect the tools that we use to be independent under a permutation of the labels. For this reason we do not use the Jaccard index as defined above for comparing network partitions; we use it instead in the paragraph on spatial overlap (at the end of the present section) to quantify the extent to which two geographical regions overlap.

The general definition of the Jaccard index for sets X and Y (as opposed to binary partitions) is given as

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}, \tag{2.6}$$

and we note in passing that /this reduces to the binary expression if we take X to be the set of nodes with label 1 according to $\mathbf{b}_X$ (and equivalently for Y and $\mathbf{b}_y$).

Other measures that generalise more easily to the situation $B > 2$ (and they do not require that the two partitions have the same number of blocks) rely on information theory. The first measure we introduce is called the mutual information. For two random variables X, Y, it is defined as the Kullback–Leibler divergence between their joint distribution and the product of their marginal distributions,

$$I(X, Y) = \sum_{x,y} \mathbb{P}(x, y) \log \frac{\mathbb{P}(x, y)}{\mathbb{P}(x)\mathbb{P}(y)}\,. \tag{2.7}$$

In other words, I quantifies the extra information needed to encode X and Y as a pair of independent random variables when they are actually dependent.

If we now consider network partitions, when we choose a node at random, the probability that its community assignment is r according to $\mathbf{b}_X$ is $\mathbb{P}(X = r) = N_r^X/N$. Similarly, $\mathbb{P}(X = r, Y = s) = N_{rs}/N$ gives the probability that a node chosen at random is in communities r and s respectively according to $\mathbf{b}_X$ and $\mathbf{b}_Y$. We use these counts to relate the community labels to the random variables X and Y and, for the problem of quantifying the overlap of two partitions,

the mutual information becomes

$$I(\mathbf{b}_X, \mathbf{b}_Y) = \sum_{r,s} N_{rs} \log \frac{N_{rs}}{N_r N_s} \,. \tag{2.8}$$

The mutual information is normally not used in this form because it has the disadvantage that the information content of a given partition $\mathbf{b}$ and all other partitions derived from it by further dividing the communities is the same and equal to the entropy of $\mathbf{b}$. Therefore, a more commonly used quantity is the normalised mutual information (NMI) [23],

$$NMI(\mathbf{b}_X, \mathbf{b}_Y) = \frac{2I(\mathbf{b}_X, \mathbf{b}_Y)}{H(\mathbf{b}_X) + H(\mathbf{b}_Y)} \,. \tag{2.9}$$

where $H(\mathbf{b}_X) = -\sum_r \frac{N_r}{N} \log \frac{N_r}{N}$ is the entropy function ($H(X) = -\sum_x \mathbb{P}(x) \log \mathbb{P}(x)$ for a random variable). The NMI is exactly 1 if the two partitions are identical, and zero if they are totally independent (this can happen if the entire network is placed inside one community).

Recently, Newman et al. [74] proposed a correction term for the mutual information to deal with an unwanted behaviour of the measure. In the situation where partition $\mathbf{b}_X$ assigns each node to a different community, the mutual information with any partition $\mathbf{b}_Y$ is maximal, even though the knowledge of the community labels in $\mathbf{b}_X$ does not give us any information about the labels in $\mathbf{b}_Y$, and the similarity should be zero. Unfortunately, the correction term in the new measure (which is called the reduced mutual information) is cumbersome to calculate.

A different but related measure that quantifies dissimilarity between partitions is the variation of information (VI) [67],

$$VI(\mathbf{b}_X, \mathbf{b}_Y) = H(\mathbf{b}_X) + H(\mathbf{b}_Y) - 2I(\mathbf{b}_X, \mathbf{b}_Y) \,. \tag{2.10}$$

The VI is a metric (in the sense of a mathematical distance function), and it is therefore useful to do a projection of the high-dimensional and discrete space of partitions into two dimensions (of the type observed in [84, 85]).

A measure of spatial overlap We have stated that one of our goals is designing a community detection algorithm that removes any spatial effects, and intuitively we expect this to result in partitions that have spatially overlapping blocks. To quantify this, we introduce a new measure that is defined for pairs of communities.

Let $\mathbf{b}$ denote a partition and let us focus on two blocks r and s. Let $R = \{\mathbf{x}_v : v \in V \,, b_v = r\}$ be the set of coordinates of the nodes in community r, and similarly define the set S for block s. Our measure is the Jaccard index calculated from the convex hulls of the two sets,

$$J_{rs} = \frac{|\operatorname{Conv}(R) \cap \operatorname{Conv}(S)|}{|\operatorname{Conv}(R) \cup \operatorname{Conv}(S)|} \,. \tag{2.11}$$

This measure satisfies $0 \leq J_{rs} \leq 1$.

2.2 Community detection

Many real-world networks are organised so that certain groups of nodes are tightly connected. Some examples include a group of friends in a social network, a functional group of proteins inside a metabolic network, or a group of web pages belonging to the same website in the world wide web [95]. These groups are what we intuitively call a *community*.

As first definition that we will revisit, a community is a cluster of nodes that are densely connected among themselves and more weakly connected to the rest of the network. *Community detection* is the problem of finding the natural division of a network in terms of communities. This problem has given rise to one of the most active fields of current research in network science (see the reviews [31, 32, 95, 103]) due to the number of applications and the different tools it supplies.

The origins of the field of community detection can be traced back to the social sciences and computer science. As early as 1955, Weiss and Jacobson studied the collaborations inside a government agency by finding work groups, considered to be "a set of individuals whose relationships were with each other and not with members of other work groups" [123]. In 1974, Hubert used graph-theoretic concepts to formalise previously proposed clustering techniques, and he applied them to the grouping of psychiatric syndromes [48].

In the field of computer science, Kernighan and Lin proposed an algorithm in 1970 to partition components of electronic circuits into two groups so as to minimise the sum of edges cut (or the sum of costs associated with edges) [52]. The field of community detection as we know it today saw the involvement of statistical physicists and applied mathematicians after the seminal contributions of Girvan and Newman [38, 75].

Communities have been used to explain the functional organisation of networked systems [38], or in the friendship, metabolic and web examples given above for which communities have a semantic interpretation. They can also shed light on network formation mechanisms, such as in the collegiate social networks studied by Traud et al. [119].

At the level of individual nodes, determining the boundary between groups can be helpful to distinguish the role of agents fully embedded inside their group or acting as mediators between different regions of the network. In his classic 1977 study, Wayne Zachary studied a university-based karate club that underwent a split [128]. He showed that he might have been able to predict the split in advance by studying the social interactions of the members using what can be understood as an early form of community detection.

In addition to this, community detection shed light on dynamical processes taking place on a network. Indeed, we can naturally expect that groups of tightly linked nodes will necessarily play a role in facilitating or obstructing the spread of opinions [104] or diseases [22].

Communities are also interesting as mesoscopic structures [95], by which we mean an intermediate level of organisation between the 'microscopic' nodes and the 'macroscopic' network as a whole. Community structure can be used for network visualization, providing a 'birds-eye' view of a large network. Furthermore, since communities represent groups with some degree of autonomy, large networks are sometimes studied by first breaking them down into smaller and

more manageable subsets of nodes [77], an approach that we follow in Chapter 4.

2.2.1 Four different perspectives of community detection

The classic view of community detection is that of assortative[1] communities: "dense subgraphs which are well separated from each other" [32]. However, community detection arises in different fields and in different use cases, and in practice the preferred definition of community is typically domain-specific [95].

Rosvall et al. [103] argue "that community detection should not be considered as a well-defined problem, but rather as an umbrella term with many facets". Since community detection is an unsupervised learning problem, they contend that the usefulness of its output depends on why we want to find community structure in the first place. This motivates a classification of the different methods according to the problem that prompted the analysis. According to this view there are four different perspectives:

1. The cut-based perspective: the defining objective is to minimise the number of links between different groups of nodes (the cut). Problems of this sort arise in computer science, for example in the form of load balancing for parallel computations or the design of computer chips (the Kernighan–Lin algorithm [52]). These application require a partition of the nodes into a predetermined number of groups of approximately balanced sizes such that the number of edges joining different groups is minimal. In this perspective there is no need for the groups to be densely connected.

2. The clustering perspective: the methods care about finding groups of with high internal edge densities [12]. This problem is closely related to minimising the cut size [63], but the main difference in this perspective is that the number of groups is typically not known a priori, and we want to find the optimal partition irrespective of the relative size of the communities and without enforcing balanced sizes. Modularity maximisation, one of the most popular community detection techniques, and also spectral methods — which make use of the eigenvalues of the Laplacian matrix [63] — are part of this perspective.

3. The stochastic block model perspective: this perspective groups nodes that are *stochastically* equivalent, i.e. connected to other nodes in the network with the same probability. This view generalises the classic assortative notion of community structure because it incorporates other types of mesoscale structure in which nodes do not have necessarily dense connections to their own group.

4. The dynamical perspective: communities are defined in terms of a dynamical process taking place on the network. While the other perspectives are essentially concerned with the structure of the network, in this perspective a community can be found as a group of nodes that trap a random walk longer than expected (Markov stability [24]), or as a

[1] Assortativity, as a node property, is the tendency of nodes to be preferentially attached to other nodes that share similar characteristics. In the context of community detection, nodes are said to belong to assortative communities when high-strength edges are more likely to be adjacent to other high-strength edges

structure that is useful to compress the dynamics of a random walk from an information-theoretic point of view (Infomap [102]).

In this work, we focus on the second and third perspectives which we now discuss in more detail.

2.2.2 Clustering data points and networks

Leaving networks aside for a moment, clustering usually refers to an unsupervised learning problem that finds structure inside a dataset. Given a collection of points in $\{\mathbf{x}_i \in \mathbb{R}^n\}$, the goal is to partition them into groups so that points in the same group are more similar to each other than they are to points inside other groups. Similarity is often written down in terms of the distance $s_{ij} = ||\mathbf{x}_i - \mathbf{x}_j||$ or a function thereof.

One of the most important tools in this context is the k-means algorithm, an iterative method to partition N data points using k Voronoi cells that divide the space around centroids (hence the name) so that within-cluster variance is minimised.

In the framework described by von Luxburg [63], a similarity matrix $\mathbf{S} = (s_{ij})$ is turned into a similarity graph which can be either fully connected, an ϵ-neighbourhood graph that connects pairs of nodes separated by a distance smaller than ϵ, or a graph that has a fixed number of nearest neighbours linked. We denote by $\mathbf{A}$ the adjacency matrix of the similarity graph, an undirected but possibly weighted network, and we define the unnormalised graph Laplacian as

$$\mathbf{L} = \mathbf{D} - \mathbf{A}\,, \tag{2.12}$$

where $\mathbf{D}$ is the diagonal matrix that holds the strength of the nodes as defined in Equation (2.2). There are other normalised versions of the Laplacian and any other one would work just as well in the present setting.

To do spectral clustering, we take the graph Laplacian $\mathbf{L}$, we compute its first k eigenvectors which are stored as column vectors inside matrix $\mathbf{U} \in \mathbb{R}^{N\times k}$, and we find the k-means partition of the N points $\{\mathbf{y}_i \in \mathbb{R}^k\}$ that are read as the rows of $\mathbf{U}$. The eigenvalues reveal more easily the clusters of points, and the k-means algorithm has an easier time with the new representation $\mathbf{y}_i$ than with the original $\mathbf{x}_i$ [63].

While we started the discussion using data points that are in Euclidean space, we described their representation in terms of a graph and it is clear that we can also employ spectral clustering if the object we want to cluster is a network. Within the clustering perspective of community detection, the clusters of "similar" data points translate into groups of nodes that have a high internal edge density, that is, an *assortative* grouping of the nodes. Next, we describe a popular quality function that belongs to the clustering perspective of community detection.

Modularity maximisation

The modularity function is a widely employed quality function that is used to compare different network partitions. Heuristically, a partition $\mathbf{b}$ is meant to score higher if the groups are dense,

i.e. if most of the edges link nodes in the same group and there are few connections that cross into different groups. Under this definition, however, the densest partition of the network is the one that has all the nodes inside a single group, so modularity compares the observed group densities against the densities that could be expected by chance. Its general form is given by

$$Q(\mathbf{b}) = \frac{1}{2M} \sum_{i,j} \big(A_{ij} - P_{ij}\big)\delta_{b_i b_j} \,, \tag{2.13}$$

where δ is the Kronecker delta and P_{ij} gives the expected number of edges between nodes i and j according to some suitable null hypothesis. Informally, the matrix $\mathbf{P}$ is called the null model. The term $1/2M$ is included for convenience so that Q measures a fraction of edges rather than an absolute number. Values of Q closer to 1 are taken to indicate a stronger community structure (although it has been pointed out that there are networks and partitions with arbitrarily high modularity that do not correspond to actual community structure [6, 44]). The matrix $\mathbf{Q} = (\mathbf{A} - \mathbf{P})/2M$ is called the modularity matrix.

Perhaps the most commonly used null model is the configuration model $P_{ij}^{\mathrm{NG}} = k_i k_j / 2M$, which is used to write the best known form of modularity, the Newman–Girvan modularity function [75],

$$Q(\mathbf{b}) = \frac{1}{2M} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2M}\right] \delta_{b_i b_j} \,. \tag{2.14}$$

The configuration model corresponds to a random reshuffling of stubs that preserves the degree distribution of the network,[2] and under this null model, modularity explicitly takes into account the heterogeneous degree distribution of real world networks.

Modularity was proposed by Newman and his collaborators in 2004. Originally, it was used to select the best partition among the candidates found using a divisive algorithm [75], but almost immediately it was recognised as a useful objective function to be maximised directly [72]. This turned out to be an extremely fertile approach and it led to a whole area of research in Network science.

A notable algorithm to maximise modularity was proposed by Clauset et al. [16]. The method improves the algorithm proposed earlier by Newman in [75], and it is known as the Fast Greedy algorithm. Starting with each node as the sole member in its community, pairs of communities are repeatedly merged if this results in the highest increase of modularity. At any step of the algorithm, each pair of communities is considered. Clauset et al. applied their method to find the community structure of a network made of more than 400,000 products purchased on Amazon, with edges connecting products that were bought frequently by the same customers. The communities obtained group books and music on similar topics (such as general interest, art, hobbies, horror, etc.), thus pointing in the direction of a semantic organization of the network.

Modularity maximisation is equivalent to an instance of the Max-Cut problem [73], which

[2] Strictly speaking, the Newman–Girvan null model preserves the degree sequence on expectation only and does not come from a true configuration model [33]. For large networks that are sparse, however, the difference becomes asymptotically small [71]. The expression employed also implicitly assumes that $M \gg 1$ because the numerator should actually read $2M - 1$.

is provably NP-hard. Heuristic approaches are thus used in almost any practical setting to find a partition that results in a local maximum. This is the case of the Fast Greedy algorithm, but also of the well known Louvain algorithm of Blondel et al. [12]. This method is closely related to the Fast Greedy algorithm, but unlike the latter, nodes are considered individually in random order, and the best increase in modularity is decided among the node's neighbouring communities (thus borrowing the Kernighan-Lin swapping steps). The candidate merges are therefore selected from a much more restricted set than for the Fast Greedy algorithm, and the resulting algorithm is significantly faster.

Modularity can also be maximised using a spectral method [73]. We use the modularity matrix $\mathbf{Q}$ which plays a role equivalent to that of the Laplacian matrix in normal spectral partitioning and we recursively partition the network, typically into two groups. For the special case of the bipartition, we use a special vector $\mathbf{s}$ — to distinguish it from the more general $\mathbf{b}$ — that has entries $s_i = \pm 1$, depending on the group assignment. Modularity is given by $Q = \mathbf{s}^\top \mathbf{Q}\, \mathbf{s}/(4m)$. The simplest way forward is to use the eigenvector associated to the leading eigenvalue to obtain a bipartition, but one can use as many eigenvectors as there are positive eigenvalues [95] (only these eigenvectors can give positive contributions to Q).

It is worth noting that modularity can be used with minimal adaptation for weighted networks simply by replacing counts of edges by sums of edge weights. The measure has also been generalised and adapted to directed networks [3], multilayer networks [68], and bipartite networks [8, 45]

Shortcomings of the modularity function

Modularity maximisation is still one of the preferred methods for doing community detection [103], even though it is well understood that the measure has several shortcomings. Most noticeably, Fortunato and Barthélemy demonstrated in [30] that standard modularity has an intrinsic resolution limit. The partition with maximal modularity balances the number of terms in the sum, each one corresponding to a community, and the value of each term. This trade-off places a limit on the number of internal edges that a group of nodes can have in order to contribute positively to Q. If the density is higher than the limit, the merging of small groups will result in a higher value of modularity.

In general, the maximisation of standard modularity is unable to find more than $O(\sqrt{M})$ communities (where M is the total number of edges). It is counter-intuitive that the detection of groups of high internal density, something that we expect to be local, is limited by a threshold that depends on the size of the whole network.

The issue of the resolution limit is partly resolved by introducing a resolution parameter γ that weights the relative importance of the null model. The resulting function is called *generalised* modularity,

$$Q(\mathbf{b};\gamma) = \frac{1}{2m}\sum_{i,j}\left[A_{ij} - \gamma\frac{k_i k_j}{2m}\right]\delta_{b_i b_j}\,. \tag{2.15}$$

The resolution parameter γ was introduced by Reichardt and Bornholdt in the context of the q-state Potts model of spin interactions [98], which can be shown to be equivalent to modularity

maximisation. The parameter γ can be interpreted as the time scale of a dynamical process [55], or motivated from the statistical fit of a particular type of stochastic block model that takes the same form as modularity maximisation [71].

Other shortcomings of modularity include the fact that large Erdős–Rényi random graphs can be shown to have partitions with high modularity without an underlying community structure [44], and that trees and tree-like networks may possess arbitrarily high values of modularity [6], even though these networks are extremely sparse. Furthermore, the modularity landscape is typically degenerate since there is a very large number of local maxima [40].

2.2.3 Stochastic block models and statistical inference

By design, methods that take the clustering perspective and maximise internal edge densities will only find assortative groups, but this view is too restrictive when it comes to describing other mesoscale structures such as disassortative communities (in which nodes are preferentially linked to nodes not in their own community), core–periphery structure [101] (core nodes are both well connected between themselves and to the peripheral nodes, while peripheral nodes are only sparsely connected between themselves) or a hierarchy of clusters [107].

The stochastic block model (SBM) is a generative network model that accommodates these more general structures. A group inside an SBM — in this context called *block* — collects nodes that are stochastically equivalent, that is, nodes that are connected to other parts of the network with similar rates and that are such that the observed differences can be attributed to random fluctuations. The model is statistical in nature and we use inference to find its parameters, particularly the partition $\mathbf{b}$ that is most compatible with our observations. A Bayesian view of statistics then says that we should be looking for the partition that has the highest probability according to

$$\mathbb{P}(\mathbf{b}|\mathbf{A}) = \frac{\mathbb{P}(\mathbf{A}|\mathbf{b})\mathbb{P}(\mathbf{b})}{\mathbb{P}(\mathbf{A})}\,. \tag{2.16}$$

The element that will occupy us in the following discussion is the likelihood function $\mathbb{P}(\mathbf{A}|\mathbf{b})$.

There has been a substantial increase of interest in SBMs due to many successful applications and theoretical guarantees about when it is possible to distinguish groups (see for example Abbe [1]). It is important to mention, however, that traditionally an SBM models the edges as being conditionally independent and this can be a poor choice for many networks that exhibit triadic closure [103] — the property in which two nodes are more likely to be connected if they share a common neighbour — though a more sophisticated model that can accommodate this property has been proposed recently by Peixoto [86].

Different formulations of the SBM

First proposed in 1983 by Holland et al. for sociometric data, the *stochastic block model* (*blockmodel* in the original) is a generative model for a network with block structure. Quoting the authors, "by block structure, we mean that the nodes of the network are partitioned into subgroups called blocks, and that the distribution of the ties between nodes is dependent on the blocks to which the nodes belong" [47].

SBMs are widely used in the field of community detection and network inference [32], and the model has accordingly been reformulated [78], extended [51, 89, 121], and re-derived from first principles [85, 88]. More than a single model, SBMs are a family of related models.

Building on the original block model, Schaub et al. [108] give a formal definition of what we call the non-degree-corrected SBM for simple graphs. It consists of a latent variable model that defines a probability over the set of binary adjacency matrices $\mathbf{A} \in \{0,1\}^{N\times N}$. Given the latent group assignment $\mathbf{b}$ of nodes, each edge A_{ij} is an independent Bernoulli random variable with a probability $\omega_{b_i b_j}$ that only depends on the group label of the nodes. We write

$$A_{ij}|\mathbf{\Omega}, \mathbf{b} \sim \text{Bernoulli}(\omega_{b_i b_j})\,, \tag{2.17}$$

where we have collected for convenience the probabilities inside the symmetric matrix $\mathbf{\Omega} = (\omega_{rs}) \in [0,1]^{B\times B}$. Peixoto [85] motivates this model as the maximum entropy ensemble of graphs for which the expected number of links between two groups $\langle e_{rs}\rangle$ is imposed.

A generalisation of the non-degree-corrected model allows multiple edges between nodes and thus gives rise to multigraphs, with elements $A_{ij} \in \mathbb{N}$. The maximum entropy ensemble gives a geometric distribution in place of Bernoulli [85], and in this model ω_{rs} represents the average number of edges between any two nodes in groups r and s instead of a probability of links being present. However, in the context of multigraphs, a much more common model considers that the number of edges between nodes follows a Poisson distribution

$$A_{ij}|\mathbf{\Omega}, \mathbf{b} \sim \text{Poisson}(\omega_{b_i b_j})\,, \tag{2.18}$$

or $A_{ii}/2 \sim \text{Poisson}(\omega_{b_i b_i}/2)$ for self-edges. This model is not a maximum entropy ensemble, but it can be seen to reduce to the geometric distribution under some assumptions [85] (one such assumption is allowing for self-edges) and it is easier to handle algebraically.

The models described above generate networks that have a Poisson degree distributions [71], which is unrealistic for most real-world networks that have instead a broad degree distributions. They also tend to group together nodes with similar degree. To address such problems, Karrer and Newman proposed the degree-corrected SBM [51]. In this model, each node is assigned a parameter θ_i that is used to control its expected degree independently of its group. The DC-SBM model becomes

$$A_{ij}|\mathbf{\Omega}, \mathbf{b}, \boldsymbol{\theta} \sim \text{Poisson}(\theta_i\theta_j\omega_{b_i b_j})\,. \tag{2.19}$$

Because the parameters ω_{rs} and θ_i always appear multiplying each other, they are arbitrary up to a multiplicative constant. The original parametrisation employed by Karrer and Newman [51] sets

$$\sum_i \theta_i \delta_{b_i r} = 1 \;\; \forall r\,. \tag{2.20}$$

With this parametrisation, the maximum likelihood estimate of the parameters gives $\hat{\omega}_{rs} = e_{rs}$ and $\hat{\theta}_i = k_i/e_{b_i}$, with e_{b_i} defined according to Equation (2.5).

Newman chooses an alternative parametrisation in [71], making $\theta_i = k_i/\sqrt{2M}$ and absorbing

the remaining constants into the ω parameters.[3] Following Pamfil et al. [81], we call ω_{rs} an edge propensity to distinguish it from the more standard degree-corrected parametrisations.

The microcanonical formulation

We often encounter more sophisticated formulations of the SBM which specify a generative processes for other aspects of the network that are incorporated by means of a hierarchy of priors and hyper-priors. For example, Newman and Reinert [76] introduce a process that generated the block memberships taking as a starting point the number of communities B and the probability γ_r that nodes belong to a particular group.

The most comprehensive and unified set of extensions corresponds to the *microcanonical* framework laid out over the past decade by Tiago Peixoto. The models that we have seen so far depend (among other things) on the propensity matrix $\mathbf{\Omega}$ which is a matrix of real numbers. It can be shown (see for example Ref [85]) that the networks generated satisfy the edge counts in Equation (2.4) only on expectation, and that the degree-corrected versions have the degree sequence $\mathbf{k}$ on expectation. These *canonical* models thus impose *soft* constraints.

In contrast to this, the *microcanonical* version the SBM depends on the matrix of edge counts $\mathbf{e}$ and the degree sequence $\mathbf{k}$ — as opposed to $\mathbf{\Omega}$ and $\boldsymbol{\theta}$ — and the ensemble of networks that can be generated satisfies the constraints exactly. The marginal likelihood is written as

$$\mathbb{P}(\mathbf{A}|\mathbf{b}) = \mathbb{P}(\mathbf{A}|\mathbf{k},\mathbf{e},\mathbf{b})\mathbb{P}(\mathbf{k}|\mathbf{e},\mathbf{b})\mathbb{P}(\mathbf{e}|\mathbf{b})\,, \tag{2.21}$$

and for suitably chosen priors, this expression has a closed-form expression. We refer the interested reader to Reference [85] for the details, and here we simply remark that the discrete random variables in the microcanonical formulation lead us to an information theoretic interpretation. Taking the logarithm of the numerator of the posterior, Equation (2.16), yields

$$\Sigma = -\log_2 \mathbb{P}(\mathbf{A}|\mathbf{k},\mathbf{e},\mathbf{b}) - \log_2 \mathbb{P}(\mathbf{k},\mathbf{e},\mathbf{b}) \tag{2.22a}$$

$$= \mathcal{S} + \mathcal{L}\,, \tag{2.22b}$$

and this quantity is called the description length. It measures the number of bits that are needed to describe the network as well as the model parameters, respectively $\mathcal{S}$ and $\mathcal{L}$.

For a suitably chosen set priors and parametrisations, the degree-corrected Poisson SBM results in exactly the same likelihood as the microcanonical one, Equation (2.21) [85, 88]. Thus, maximising the posterior distribution $\mathbb{P}(\mathbf{b}|\mathbf{A})$ is equivalent to minimising the description length. The inclusion of the priors prevents the model from overfitting the data because a large number of blocks will result in a smaller $\mathcal{S}$ but a larger $\mathcal{L}$, and the parameters that best describe the data will balance the two terms. Searching for the point estimate that minimises Σ is called the minimum description length (MDL) principle.

Peixoto has also extended the microcanonical SBM to weighted [89] and layered networks [87]. This later use case will be employed in Chapter 4 so we describe it next.

[3] Strictly speaking, Newman's formulation does not include the degree sequence as a parameter; he is writing instead the profile likelihood $\mathbb{P}_{\hat{\theta}}(A|\mathbf{\Omega},\mathbf{b})$ in which $\boldsymbol{\theta}$ is replaced by its maximum likelihood estimate.

Suppose that a network has C layers. In the *edge covariates* model, the collapsed graph $\mathbf{A} = \sum_l \mathbf{A}^l$ is generated according to an SBM with the nodes partitioned into $\mathbf{b}$, and the edges are distributed randomly across layers conditioned on the layer edge counts e^l_{rs} (the collapsed edge counts are $e_{rs} = \sum_l e^l_{rs}$). Importantly, if we place ourselves in the degree-corrected framework, the degrees are constrained in the collapsed graph and the edges incident to a node will be distributed randomly across layers.

An alternative to this is a model with *independent layers* that are each generated by separate SBMs that are only constrained in that the block membership is the same for all the layers. This model works particularly well with degree-correction because each layer has the edge counts and degree sequence as parameters (if the degree-correction is omitted there is an extra binary variable indicating whether node i can have incident edges in layer l). Furthermore, for the degree-corrected version of the SBM, the independent layers model encompasses the edge covariates model in that there are networks that can be generated with the former that will only be sampled with vanishingly small probability with the latter; on the other hand, all the networks that can be generated with the edge covariates can also be sampled with independent layers.[4]

Methods for statistical inference

There are a number of alternative approaches that are used to infer the parameters in the statistical models presented above. Typically, we search for maximum likelihood or maximum a posteriori estimates. In any realistic situation, the exact inference is intractable and we need tools to find local maximisers that approximate the estimates.

An important such tool is the expectation-maximisation (EM) algorithm, introduced in a general form by Dempster et al. [25]. The EM algorithm is used when the computation of the marginal likelihood function is intractable, for example due to missing data. The EM algorithm alternates between an expectation step — in which we use the current estimate of the general set of model parameters, $\boldsymbol{\theta}^{(t)}$, to calculate the expected value of the log-likelihood function — and a maximisation step — during which we find $\boldsymbol{\theta}^{(t+1)}$, the updated set of parameters that maximise the previous expected value. According to Mariadassou et al. [64], the EM algorithm cannot be applied directly to network data and therefore they propose a variational method that maximises a lower bound to the log-likelihood function.

A simpler method grounded on the modelling framework where the group assignments are parameters, not missing data, was introduced by Karrer and Newman for the maximisation of the degree-corrected log-likelihood [51]. Nodes are initially divided at random into B groups (the number is selected *a priori*) and subsequently nodes are moved to the group that will most increase the objective function (or least decrease it), with each node being moved once. After a full sweep of all the nodes, the state with the highest likelihood is used as the initial partition in another optimisation round, and the process is repeated until a full sweep does not result in an increase.

Perhaps the most powerful methods are Markov Chain Monte Carlo (MCMC) methods. They

[4]Without the degree-correction, the edge covariates and independent layers models are asymptotically equivalent if the edge counts in each layer are large enough [87].

were employed in an early effort by Nowicki and Snijders [78] in the form of a Gibbs sampler to infer the parameters of the SBM without degree-correction. More recently, Newman and Reinert [76] use MCMC importance sampling in order to approximate the marginal distribution for the number of communities b. The space defined by the parameters $(\mathbf{b}, B)$, with associated probability $\mathbb{P}(\mathbf{b}, B|A)$, is explored using separate schemes to sample over $\mathbf{b}$, holding B fixed, and vice-versa to sample over B, holding $\mathbf{b}$ fixed (crucially, empty groups have to be allowed).

Riolo et al. [100] propose a more efficient MCMC algorithm in which communities are always non-empty and nodes are moved either to an existing group (if the node is the sole member of its community this decreases B) or to a new one (this increases B). The probability $\mathbb{P}(\mathbf{b}', B'|\mathbf{b}, B)$ of moving from state $(\mathbf{b}, B)$ to state $(\mathbf{b}', B')$ — decomposed as the probability of proposing the movement and the probability of accepting it — and the probability of returning are required to satisfy the condition of *detailed balance*,

$$\frac{\mathbb{P}(\mathbf{b}', B'|\mathbf{b}, b)}{\mathbb{P}(\mathbf{b}, B|\mathbf{b}', B')} = \frac{\mathbb{P}(\mathbf{b}', B'|\mathbf{A})}{\mathbb{P}(\mathbf{b}, B|\mathbf{A})}\,, \tag{2.23}$$

and the acceptance probability is chosen accordingly. The move proposal selects existing groups r and s uniformly at random and proposes to move a node from r to s, or proposes to create a new group.

It is worth noting that the new MCMC algorithm, while equivalent to the earlier version by Newman and Reinert, is based on a different prior, $\mathbb{P}(\mathbf{b}, B)$. In [76], the authors use uniform priors $\mathbb{P}(B) = 1/N$ (which in practice assigns too much importance a large number of communities) and $\mathbb{P}(\boldsymbol{\gamma}|k)$. In [100], the prior is implicitly defined by a generative model in which the nodes are sorted in random order, and each one is either placed in the previous node's group with probability $1 - q$ or in a new group with probability q. In the large N limit and for a suitably chosen q, this prior becomes $\mathbb{P}(k) \propto 1/(k-1)!$.

In the context of the microcanonical SBM and the minimum description length principle, Peixoto [85, 90] also uses an MCMC algorithm to sample the marginal posterior distribution $\mathbb{P}(\mathbf{b}|\mathbf{A})$, adopting the Metropolis–Hastings acceptance criterion to ensure detailed balance. Furthermore, Peixoto adapts the method of simulated annealing — where $\mathbb{P}(\mathbf{b}|\mathbf{A})$ is replaced by $\mathbb{P}(\mathbf{b}|\mathbf{A})^{\beta}$ and β is progressively increased towards $\beta \to \infty$ — in order to maximise the posterior distribution.

In both the sampling and the optimisation, Peixoto introduces a heuristic choice for the move proposals, using the current partition $\mathbf{b}$ and the selected node's neighbourhood in order to guide its next movement. The details can be found in [85]; we simply note that ergodicity is still assured and the method is not biased towards a specific mixing pattern, specifically not an assortative one (unlike the Louvain algorithm that proposes movements only to the neighbouring communities).

2.3 A first analysis of the retail data

We now provide a first exploration of the retail data and we present the most obvious (but ultimately unsuccessful) way of modelling the data as a bipartite network that we call the customer-store network.

Network models have been applied before to study the retail market, in particular to examine customer-product networks. For example, Pennacchioli et al. [92, 93] relate the volume of product sales with the set of customers buying them, and give a relationship in terms of product complexity. Closer to our interests, Pamfil [82] investigates the community structure and applies the findings to the problems of customer segmentation and product recommendation.

Through our industrial partner, Dunnhumby, we have access to four year's worth of anonymised shopping records inside a large UK retail chain and covering the fiscal years 2017 to 2020. However, we were fist given the data for 2017 and only when we started the analyses that are presented in Chapter 5 were we able to extract the remaining years. So let us now introduce the data for the first year.

The data was generated by over 1.4 million loyal households shopping for groceries across $3,149$ stores. Individual customers are identified by their loyalty card, but different cards — typically only one or two, in very few instances a few more than that — can be joined into one household. Here, we pay no attention to this distinction and we use the terms *household* and *customer* interchangeably.

The data is pseudonymised with customers identified using a numeric household ID. We have been given the geographical coordinates of a majority of the stores, but we were neither given nor had a way of inferring the location or home address of customers.

The records for 2017 were initially aggregated in time (this changes in future chapters): for each store visited we have access to the total number of times that household visited, the total spend, the total number of items and the total number of unique items bought, over the whole year. We are also given the format of the store — Convenience, Supermarket, Petrol, Hypermarket and High St — and we are told if the shopping was done online or offline.

2.3.1 Loyal customers

The loyal customers from which our data is sampled are customers that shopped regularly throughout the year. Specifically, loyal customers are those placed inside the loyal segmentation for at least 46 weeks and with a minimum of 36 visits during the year.

In Figure 2.2a, we plot the relative frequency of yearly spend. We see that the distribution is approximately log-normal and it has a pronounced peak. While the spread between the minimum and maximum spend covers several orders of magnitude, we find a majority of the customers in close proximity to the peak. Indeed, the spend of 70% of all households falls within one geometric standard deviation of the geometric mean (respectively e^{σ} and e^{μ}, where μ and σ are the mean and standard deviation of the logarithm of the values).

In Figure 2.2b, we show the relative frequency of yearly spend for the online channel only. We note that the peak is lost because of the behaviour of customers that buy mostly in physical

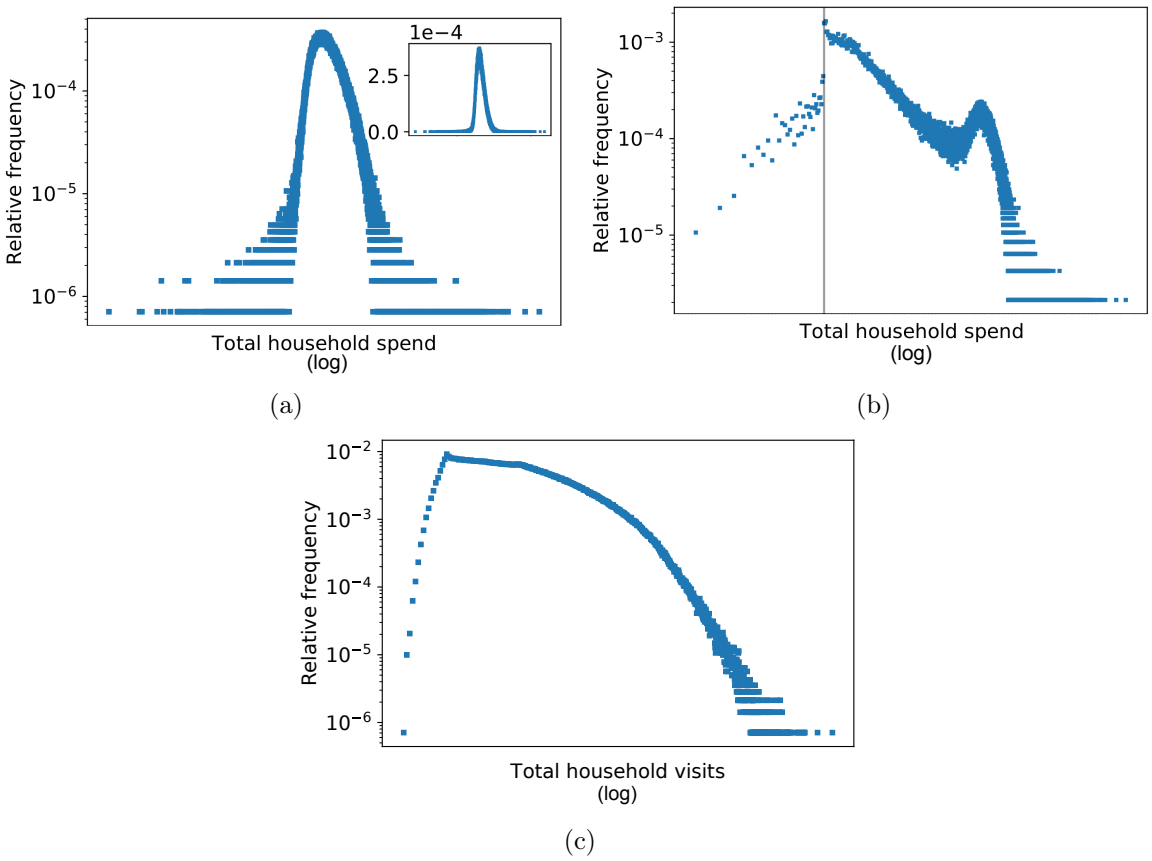


Figure 2.2: (a) Relative frequency of yearly spend by 1.4 million households (rounded to the nearest integer) and (b) computed only from online shopping. We note that the minimum spend for free delivery of online orders is £40 (indicated by the vertical line). (c) Relative frequency of the number of yearly visits by household. For anonymisation reasons, the values of the x axes are not shown.

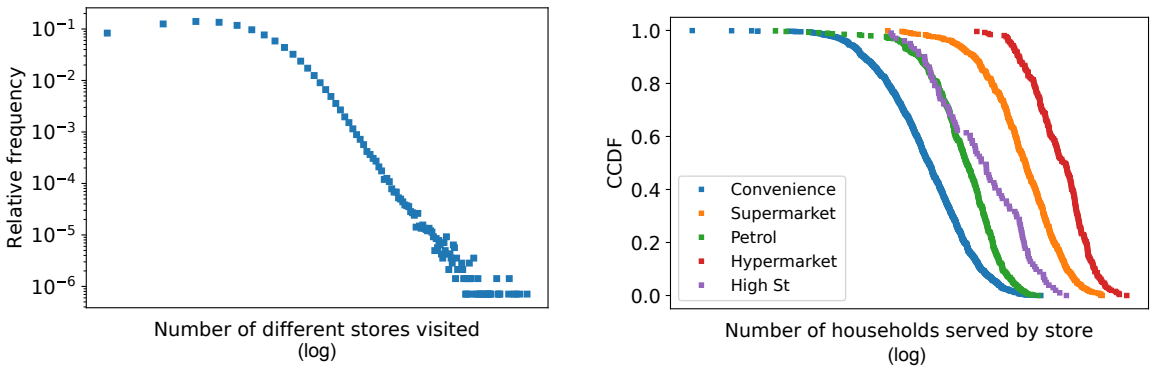
stores and only shop online a few times during the year. From the point of view of what defines loyal customers, this means that it is important to consider online and offline shopping together. Furthermore, the minimum spend to qualify for free delivery is indicated with a vertical line and we can see that this threshold makes the low spend shops significantly less common, something we do not see for the shopping done offline.

Unsurprisingly, the total household spend is highly correlated with the yearly number of items bought (Pearson correlation of $\rho = 0.846$) and also positively correlated with the number of visits ($\rho = 0.231$). The distribution of the latter quantity can be seen in Figure 2.2c.

2.3.2 Online shopping

We have seen an example of the distinction between online and in-store shopping and we encounter many more particularities that are relevant for roughly one third of all the customers that shopped online at least once during 2017.

Online baskets tend to be larger than in store baskets, accounting for around 7% of total visits (i.e. individual baskets) but approximately 20% of the total spend.



(a) (b)

Figure 2.3: (a) Distribution of the number of stores visited by each household and (b) complementary cumulative distribution function (CCDF) of the number of households served by each store. For anonymisation reasons, the values of the x axes are not shown.

When a customer shops a basket online, their order is served from a physical store[5] which is recorded in our data. The store is chosen by the *Retailer*, probably based on the availability of stock and the distance to the delivery address (although we have no precise information on this) and the customer is unaware of which store is selected. In this sense, the relationship between the customer and the physical store that served their order is artificial since it does not result from a decision made by the customer.

This raises the question of how to treat the relationship between customers and stores in the context of online shopping. One option is to discard the information about the physical store that served the orders and treat all of online shopping as originating from the special 'Online' store, but in this case we must deal with this store being connected to customers all across the country.

Alternatively, we can disregard the channel information and treat the online visits and spend as actually originating at the physical stores. This would be more suited to do a network projection where we connect two stores that share the same customers. However, it has the obvious disadvantage that it disregards the lack of choice made by shoppers.

2.3.3 Degree distributions

We characterise what are the most common behaviours in terms of the number of stores that the customers visit (Fig. 2.3a) as well as the number of customers that the stores serve (Fig. 2.3b). In network terms, these are the degree distributions of the two types of nodes defined by the bipartite data. For the purpose of these calculations, we treat online shopping as a single store and the visits for the different formats are only counted for customers that physically visited the stores (i.e. we are in the setting of the first solution to the problem of online shopping proposed above).

In Figure 2.3a we can see that a majority of customers visit only a handful of different

[5]More precisely, the order is served from a store that belongs to one of the larger formats: Hypermarket and Supermarket.

stores in a normal year, and the probabilities for the first few numbers are approximately equal. On the other hand, the probability of finding a customer that visits more stores drops quickly. Surprisingly though, a handful of customers have shopped in hundreds of stores.

The finding above suggests that there are two different shopping behaviours. The left hand side of the plot tells us that customers that shop in a handful of stores are, to a first approximation, found in equal numbers. This is probably a result of local visits to nearby stores and we can easily image that any household that frequents, for example, three stores could easily visit a fourth one. On the other hand, the right hand side of the plot tells us about a second behaviour that corresponds to customers that travel extensively. We imagine that in this setting visiting one more store requires a substantial extra effort, and this probably explains the sharp decrease in the frequency.

From the perspective of the stores, we give the complementary cumulative distributions of customers served in Figure 2.3b. We conclude that there is a clear ordering based on the size of the store type. Interestingly, we can see a transition for the format High St: the smaller shops in this format serve about the same number of customers as the Petrol format, but the larger stores serve many more.

2.3.4 Bipartite network construction

We use the shopping records to build a bipartite network that links customers to the stores they visit. We denote by $\mathcal{H}$ and $\mathcal{S}$ the set of households and stores respectively, and the set of edges is $E \subset \mathcal{H} \times \mathcal{S}$. An edge $e = (h, s) : \ h \in \mathcal{H}, \ s \in \mathcal{S}$, has associated weight w_{hs}.

We can use different sources of information for the edge weights. The variable that we found to be the most suitable, both from a modelling perspective and from the fact that it varies in a more limited range (see Figure 2.2), is the number of visits. Therefore, we let w_{hs} denote the number of visits that household h made to store s. We use the biadjacency matrix $\mathbf{B} = (B_{hs})$ of size $|\mathcal{H}| \times |\mathcal{S}|$ to represent our network. The entries satisfy $B_{hs} = w_{hs}$ whenever $(h, s) \in E$, and $B_{hs} = 0$ otherwise.

We have touched before on the issue of online shopping and we said that we can build a network in which online is an individual node, or we can drop the distinction of the channels altogether. These two modelling options will define two different networks denoted G^{OS} (for online store) and G^{NoC} (for no channel). Discarding the totality of online shopping defined results in the network G^{Off} (for offline), that we use for comparison only.

2.3.5 A first attempt of detecting communities

The Infomap algorithm which minimises the objective function called the map equation — proposed by Rosvall et al. [102] — can be applied directly to bipartite networks. We use it to find partitions in our customer-store networks to explore what is the best way to simplify them. By employing the map equation, we are placed in the dynamical perspective of community detection and communities are groups of nodes that trap a random walk for a significant amount of time.

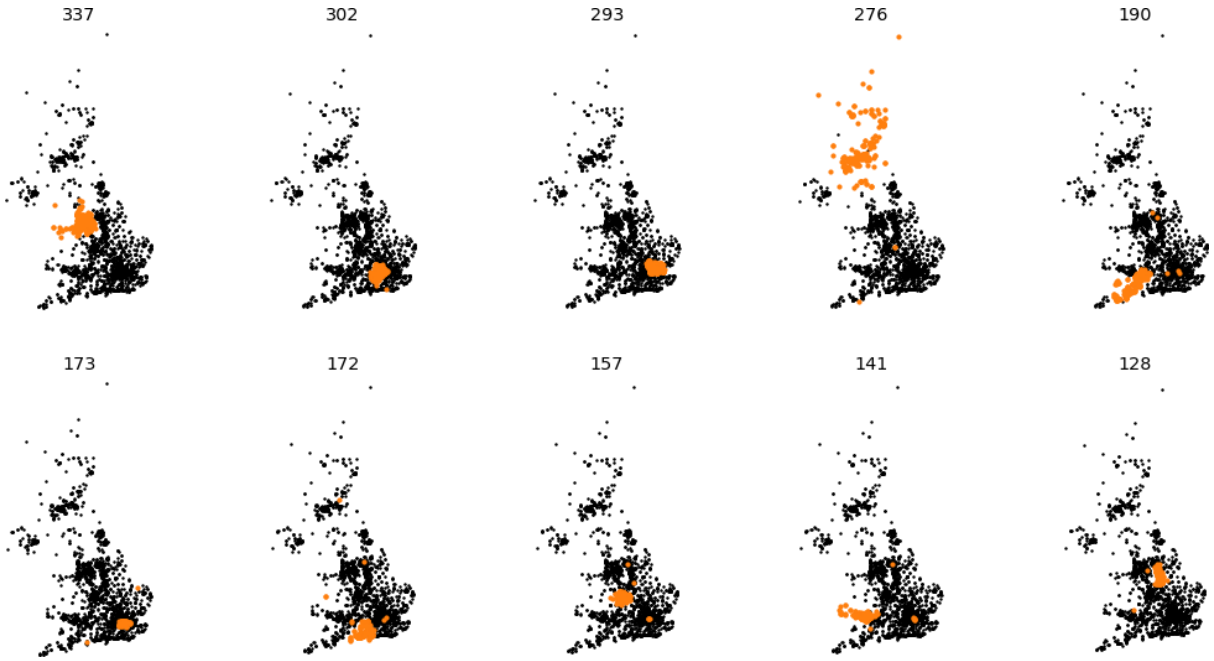


Figure 2.4: Representation of the ten largest communities measured by the number of stores (the number at the top of each plot). The stores that belong to the community are highlighted in orange (the customers assigned to each community are not shown). In total there are 31 communities.

We apply the algorithm to the networks G^{Off}, G^{OS} and G^{NoC}. It is not much of a surprise that the partition resulting for the offline information only, G^{Off}, is the most fragmented, with 36 different communities that have on average 87 stores in them. The network that does not distinguish between the online and offline channels, G^{NoC}, is partitioned into 31 communities with an average of 102 stores in them, and the partition found for G^{OS} is composed of 30 communities with an average of 105 stores in them. However, the similarity in these numbers hides that Infomap finds a giant community G^{OS} with almost 60% of all the stores in it.

We can quantify the distance between partitions using the NMI. We find that the agreement between the partitions of G^{Off} and G^{NoC} is 0.863, and this is a larger agreement than the NMI between G^{NoC} and G^{OS} at 0.735. In other words, the amount of connectivity introduced in G^{OS} and the giant community found result in a partition that is unlike what we see for the other two networks.

Using this comparison as a guide, it would seem that the most useful partition has 31 communities and it was obtained for G^{NoC}. Even though the online relationship between households and stores can be deemed artificial, the result is not too far from what is achieved by discarding all of the online information, and it has the advantage of being less fragmented.

We find a representation of the stores in the ten largest communities in Figure 2.4. Stores placed inside a community are coloured in orange. The communities found are strongly related to geographical regions, a result that is not surprising since distance must be a key driver of store selection and stores in close proximity are more likely to share customers. However, the mostly geographical communities are not necessarily the most useful way of breaking up the network because it does not add great amounts of new information. We need to take into account

that our system is embedded in space to find a more interesting mesoscale organisation that is "space-corrected". Developing such a methodology will be the topic of the next two chapters.

Chapter 3

A constrained framework for models of human mobility

In this chapter we discuss two models of population-level mobility called the gravity law and the radiation model. The gravity law is popular with different disciplines and this fact has led to an increasingly confusing literature with a multiplication of models. Below, we present an opinionated framework that supersedes many of these models and allows us to interpret the gravity law in terms of a family of statistical models which are crucial to the "space-corrected" SBM methodology that is developed in the following chapter.

3.1 Aggregate trip distribution models

There are two main modelling paradigms into which most population-level mobility models fall: gravity models and intervening-opportunities (IO). Gravity models consider that the flow between two locations decays with the distance separating them due to an increasing cost of travel.[1] For its part, inside an IO model, the flow between locations decreases if there is a higher number of locations other than the destination that are situated closer to the origin [117]. Distance plays an indirect role inside the IO models in that the further two locations are, the more likely it is that there are alternative destinations that travellers could choose for convenience. The radiation model belongs to the IO family [111].

3.1.1 The origin-destination matrix

Mobility models, also called trip distribution models, predict the number of trips between a *discrete* set of N locations or zones. Inside a city, these could correspond to the cells in a rectangular grid or the cells defined by mobile phone towers; at a larger scale we can take administrative divisions or even cities. If the data itself does not define the discretisation of the space, it is left to the practitioner to find the right level of coarse-graining and this is generally a challenging task.

[1] Here we take the word flow to mean the quantity measuring the spatial interactions being modelled. Often, this will be the number of trips between locations, and in this work we use the terms interchangeably.

The object of study is the $N \times N$ matrix $\mathbf{T} = (T_{ij})$, which records the number of trips that are made between pairs of location over a predetermined period of time. This matrix is called the *origin-destination* (OD) matrix. Since the counting depends on the time window that is chosen — 5,000 trips between two locations have a different meaning depending on whether they take place over an hour, a day or a year — we see that T_{ij} naturally comes equipped with physical dimensions: number of individuals (which arguably can be thought as dimensionless) per unit time. We use $\mathbf{T} = (T_{ij})$ for an empirically observed OD matrix, and below we refer to a model-derived matrix using $\mathbf{T}^* = (T^*_{ij})$ where we take the star symbol to indicate a generic model, or we employ specific letters to designate a particular model.

The data recorded inside an empirical OD matrix might or might not be stationary. For example, in the census data that we analyse below, participants are asked to name the local authorities in which they reside and work, and as such, the data represents a snapshot in time which is stationary by definition. (A separate question would be studying the change between different censuses.) The OD matrix derived from shopping data, on the other hand, corresponds to aggregated trips obtained over the course of a year. In order to model this data using mobility models, we have to assume that we have a sufficiently large number of customers, and that observing them for a year is a sufficiently long period of time to obtain stationary shopping patterns.

3.1.2 Inputs used in the models

Mobility models always use as an input the coordinates of the locations $\{\mathbf{x}_i \in \mathbb{R}^2\}_{i=1,\dots,N}$ (sometimes indirectly via the matrix of pairwise distances). In addition to this, the models measure the relative importance of the locations using a variety of different inputs, for example the population at each location, their geographical extension, or the number of trips entering or leaving each unit. To compare the predictions made with different models, it is important to do it with respect to the same input [58]. In our work we take the approach of selecting the number of outgoing and incoming trips as inputs and these can be derived empirically from $\mathbf{T}$. We store the $2N$ variables inside the vectors $\mathbf{O}, \mathbf{D} \in \mathbb{R}^N$. In most circumstances, we will refer to specific entries and write, for example, O_i for the total flow leaving the origin i and D_j for the total flow arriving at destination j.

We will assume that the systems we study are closed: all of the trips that originate among the N locations also arrive at one of these locations. This condition implies that $\sum_i O_i = \sum_j D_j = M$, the total number of trips. The reason for making this assumption is that below we introduce an iterative algorithm whose convergence can only be guaranteed for closed systems. However, the algorithm is only used with some of the models and it would still be possible to work with systems that are not closed.

In our work, we follow the convention that all the trips recorded in $\mathbf{T}$ take place between *different* locations, and we set $T_{ii} = 0$ for all i. This is common for studies involving the gravity model because its predictions are degenerate when the distance between locations is zero. There are ways around this and some studies select the diameter of the Voronoi cell around each location as the representative distance from a node to itself. Still, the main reason that we have

for discarding self-edges is that our use for human mobility models has to do primarily with an application to find groups of interacting locations (which we will cover in Chapter 4) and including self-edges seems to lead to more fragmented groups of nodes.

For our models, this leaves us with $N(N-1)$ unknowns. Assuming that the space is two-dimensional, we specify the positions of the different locations with $2N$ input variables; in addition to this, the variables O_i and D_j also count as $2N$ inputs. Therefore, as long as $N > 5$, there are more unknowns than inputs.

3.1.3 Constrained models

Despite claims that humans or other goods move in a predictable an universal way, real-world systems are incredibly noisy and the best a model can do is predict the expected value for the flow between locations which we denote $\langle T^*_{ij} \rangle$. In other words, we argue that mobility models should be statistical in nature and that we should treat T^*_{ij} as a random variable.

In the literature about human mobility, this is sometimes implicitly assumed but stating as much explicitly — like Simini et al. do for their radiation model [111] — is not common practice (and even they claim, incorrectly, that the gravity law is deterministic).

We propose a framework that is particularly suited for writing the gravity law and other types of trip-distribution models as statistical models, which is based on constrained models. This family of models was introduced by Wilson in 1971 [124] and there are three types of constrained models in it.

In the *production*-constrained case, the model satisfies

$$\left\langle \sum_j T^*_{ij} \right\rangle = \sum_j \langle T^*_{ij} \rangle = O_i \;\; \forall i \,. \tag{3.1}$$

Alternatively, an *attraction*-constrained model satisfies

$$\left\langle \sum_i T^*_{ij} \right\rangle = \sum_i \langle T^*_{ij} \rangle = D_j \;\; \forall j \,. \tag{3.2}$$

If both (3.1) and (3.2) hold simultaneously, the model is *doubly*-constrained (also called by Wilson *production-attraction* constrained).

It is worth emphasising that, in the two equations above, the right hand side is the data. The flows O_i and D_j are obtained by summing the rows and the columns of the empirical OD matrix $\mathbf{T}$, but that is not the meaning of the equations above; instead Equations (3.1) and (3.2) define the constraints on the model.

Constrained models are used frequently but they are not often recognised as such: the radiation model of Simini el al. [111] and the distance null model of Liu et al. [61] are two examples of production constrained models.

3.1.4 Leveraging the constraints for statistical models

Suppose that we use a trip-distribution which predicts an interaction of size $t_{ij}(\mathbf{x}_i, \mathbf{x}_j; \boldsymbol{\theta})$ (or simply t_{ij} for notational convenience) for a given pair of locations by means of their coordinates and possibly a set of parameters $\boldsymbol{\theta}$. We name this the *affinity* function, and there will be as many types of functions as there are flavours of trip-distribution models.

Production constraints: We can write a model that satisfies Equation (3.1) in the following way: we first normalise the interactions to obtain a row-stochastic matrix and then we draw the i-th row of matrix $\mathbf{T}^*$ from a multinomial distribution

$$p_{ij}^{\mathrm{PC}} = \frac{t_{ij}}{\sum_s t_{is}}\,, \tag{3.3a}$$

$$(T_{ik}^{\mathrm{PC}})_{k=1,\dots,N} \sim \mathrm{Mult}\left(O_i, (p_{ik}^{\mathrm{PC}})_{k=1,\dots,N}\right)\,, \tag{3.3b}$$

where the probability mass function of the multinomial distribution is

$$f(x_i, \dots, x_r; N, p_1, \dots p_r) = \prod_k \frac{N!}{x_k!} {p_k}^{x_k}\,. \tag{3.4}$$

The univariate marginal for the i-row and j-th column is binomially distributed,

$$T_{ij}^{\mathrm{PC}} \sim \mathrm{Bin}(O_i, p_{ij})\,, \tag{3.5}$$

and its expected value is

$$\left\langle T_{ij}^{\mathrm{PC}} \right\rangle = O_i\, p_{ij}^{\mathrm{PC}}\,. \tag{3.6}$$

Clearly, this statistical model satisfies Equation (3.1). We note that for this model to make sense, we require $O_i \in \mathbb{N}$.

Attraction constraints: In an analogous way, the attraction-constrained case is based on a column-stochastic matrix with the random vector corresponding to the j-th column of the attraction-constrained OD matrix distributed according to

$$p_{ij}^{\mathrm{AC}} = \frac{t_{ij}}{\sum_s t_{sj}}\,, \tag{3.7a}$$

$$(T_{kj}^{\mathrm{AC}})_{k=1,\dots,N} \sim \mathrm{Mult}\left(D_k, (p_{kj})_{k=1,\dots,N}\right)\,. \tag{3.7b}$$

The univariate marginal is $T_{ij}^{\mathrm{AC}} \sim \mathrm{Binom}(D_j, p_{ij})$ and the expected value is $\langle T_{ij}^{\mathrm{AC}} \rangle = p_{ij}\, D_j$. Again, we must ask that $D_k \in \mathbb{N}$.

Double constraints: Following Wilson, we write the doubly-constrained model as

$$\langle T_{ij}^{\mathrm{DC}} \rangle = A_i O_i\, t_{ij}\, B_j D_j\,, \tag{3.8}$$

where the terms A_i and B_j are called balancing factors and they ensure that the constraints (3.1) and (3.2) are satisfied. There is no closed-form expression for them, but substituting the previous expression into (3.1) and (3.2) yields

$$A_i = \frac{1}{\sum_j t_{ij}(B_j D_j)}\,, \qquad\qquad B_j = \frac{1}{\sum_i (A_i O_i) t_{ij}}\,. \tag{3.9}$$

as long as $O_i, D_j \neq 0 \,\forall\, i, j$. These two equations suggest that we solve iteratively for A_i and B_j using, for example, the starting guess $B_j = 1$. This is called the iterative proportional fitting procedure (IPFT) [28]. Necessary conditions for this approach to converge for a non-negative matrix (t_{ij}) are that $O_i, D_j > 0$ (which we already required) and that (t_{ij}) has nonzero support for all the rows and columns.[2]

Whenever the method converges, we factor the doubly-constrained model as

$$\langle T_{ij}^{\mathrm{DC}} \rangle = O_i \, p_{ij}^{\mathrm{DC}}\,, \tag{3.10}$$

where $p_{ij}^{\mathrm{DC}} = A_i t_{ij} B_j D_j$ defines a row-stochastic matrix (the alternative is to factor the D_j term and get a column-stochastic matrix). With this choice, the production constraint will be satisfied exactly and the remaining attraction constraint will be satisfied on expectation.

The factorisation of Equation (3.10) is the missing ingredient we were missing to interpret the doubly-constrained T_{ij}^{DC} as a random variable. We use this interpretation extensively in the next chapter to compute what we call the spatial backbone, a signed network where only the flows that are statistically significant are preserved.

3.2 The gravity law

The gravity law is one of the most widely used modelling tools in the field of transportation planning. It is used for example to model commuting flows [7], cargo ship trade flows [50], and volumes of inter-city phone calls [54].

The gravity law is used inside a family of closely related models often expressed using different notations. We have put great emphasis in selecting a consistent notation in what follows. The introduction of the affinity function allows us to easily derive the different models in the family and to clarify which parameters are being used.

The starting point of the gravity law is writing down the volume of interaction between two locations separated by a distance d_{ij} as

$$t_{ij} = m_i n_j \, f(d_{ij})\,, \tag{3.11}$$

where m_i and n_j reflect the relative importance of the origin and the destination, respectively. Geographers sometimes dub these quantities the propulsiveness of the origin and of the attractiveness of the destination [34]. These two variables need not be the same which results in a

[2] Sufficient conditions are harder to state. The main result goes by the name of Sinkhorn's Theorem and we refer the interested reader to references [13, 112, 113].

directed model in which case $T^*_{ij} \neq T^*_{ji}$ in general. The function f is called the *deterrence* function, and normally it will be a decreasing function of its argument. A common choice for the deterrence function is

$$f(d) = d^{-\gamma}\,, \tag{3.12}$$

and this form, which is reminiscent of the law of universal gravitation, gives the model its name. Another option is an negative exponential $f(d) = \mathrm{e}^{-d/\ell}$, used for example by Cerina et al. [14]. We note that we can write $d^{-\gamma} = \exp(-\gamma \log d)$, and this makes it explicit that the power law function is not a different deterrence function to the negative exponential, it is instead a re-scaling of the (possibly generalised) distance which is measured on a logarithmic scale. In our work, we mostly focus on the power law deterrence function, Equation (3.12), but the arguments we will make can be generalised to other one-parameter functions.

Besides the function f, there is a great variety of choices that are made in the literature regarding the nature of the features m_i and n_j. In Ref. [130], Zipf postulated in 1946 that the movement of persons between cities is proportional to the product of the populations P_iP_j, and since then the population at each location is generally the most important quantity associated with the gravity law. A more modern form is common, introduced by Alonso [125], that raises the population to a power and thus introduces additional parameters: $m_i = {P_i}^{\alpha}$ and $n_j = {P_j}^{\beta}$. Other modelling choices include the number of job openings at the destination, the retail surface of a mall, the GDP of a country, or other features that would be correlated with the flow between locations. This raises the following question: are there canonical variables to measure the propulsiveness and attractiveness of the locations of interest?

A far less common choice is the empirical outgoing and incoming flows, but here we argue that these are as the canonical variables. We use $m_i = {O_i}^{\alpha}$ and $n_j = {D_j}^{\beta}$. We note that in the context of a weighted network, O_i and D_j correspond to the row and column sum of the weighted adjacency matrix and therefore they are in effect the out- and in-strengths respectively. Masucci el al. [65] employ the empirical flows but only inside the radiation model, and Ying et al. [127] rely on them to define the gravity model but they set $\alpha = \beta = 1$. It seems that, in a mobility study, only Piovani et al. [94] use them in the same way we do, but their work only focuses on the doubly constrained model.[3] In a different context to mobility, Dong et al. make use of the gravity formula with the empirical flows [26].

Putting all the ingredients together, the version of the affinity function that we propose is

$${t_{ij}}^{\mathrm{Grav}} = \frac{{O_i}^{\alpha}{D_j}^{\beta}}{{d_{ij}}^{\gamma}}\,. \tag{3.13}$$

3.2.1 The unconstrained model

The simplest way obtain the expected flow from the affinity function (3.13) is by means of the unconstrained model,

$$\langle T^{\mathrm{UC}}_{ij}\rangle = K\,{t_{ij}}^{\mathrm{Grav}}\,, \tag{3.14}$$

[3] Piovani et al. actually describe their model in terms of the population P_i at the origin and the employment E_j at the destination but, given how they define these quantities, we conclude that they are actually employing the empirical flows.

where K is a constant of proportionality which is computed using the total number of trips $M = \sum_i O_i$ as

$$K = \frac{M}{\sum_{r,s} {t_{rs}}^{\text{Grav}}} \,. \tag{3.15}$$

This makes sure that the total flow predicted by our model matches the one that is empirically observed.

The constant K has dimensions so a to ensure that the dimensions of the left and right hand side of the equation match, and we can see they will depend on the value of the parameters α, β and γ. Unfortunately, this means that in general K will not have the same dimensions for two different datasets. Moreover, if we change the features used to measure the propulsiveness or the attractiveness of the locations, the dimensions of K will also change. We have to conclude that the unconstrained gravity model does not deal with physical dimensions satisfactorily.

3.2.2 The production-constrained model

Substituting the affinity function ${t_{ij}}^{\text{Grav}}$ into the production-constrained probability (3.3a), we note that the term ${O_i}^{\alpha}$ factorises out of the sum in the denominator and gets cancelled. More generally, we can multiply each row inside the affinity matrix by a constant without affecting the prediction of a production-constrained model.

It follows that we can specify the production-constrained gravity model using only the parameters $\boldsymbol{\theta}^{\text{PC}} = (\beta, \gamma)$. The prediction for the expected flow between locations i and j will be given by

$$p_{ij}^{\text{PC}} = \frac{{D_j}^{\beta} / {d_{ij}}^{\gamma}}{\sum_s {D_s}^{\beta} / {d_{is}}^{\gamma}} \,, \tag{3.16a}$$

$$\langle T_{ij}^{\text{PC}} \rangle = O_i \, p_{ij}^{\text{PC}} \,. \tag{3.16b}$$

We note, for future comparison with the attraction-constrained model, that the empirical flows play complementary roles with the incoming trips D_j appearing inside the probabilities and the outgoing trips O_i ensuring that the constraints are satisfied.

Since the term p_{ij}^{PC} quantifies the probability that a trip that we know started from i will go to j, the production-constrained model is the basis of a very simple model of individual mobility based on random walks. However, such a model will only be useful in a subset of real-world scenarios because it will make a location's outflow equal to its inflow (i.e. individuals are conserved).

3.2.3 The attraction-constrained model

This time, we can multiply each column of the affinity matrix by a constant term without changing the final expression, and thus the term ${D_j}^{\beta}$ is irrelevant for the final prediction of the constrained model and we specify the attraction-constrained gravity model using the two

Constraint type	Relevant params.	Affinity function	Prediction (exp. value)
Unconstrained	(α, β, γ)	$t_{ij} = {O_i}^{\alpha} {D_j}^{\beta} \, {d_{ij}}^{-\gamma}$	$T_{ij}^{\mathrm{UC}} = K t_{ij}$
Production	(β, γ)	$t_{ij} = {D_j}^{\beta} \, {d_{ij}}^{-\gamma}$	$T_{ij}^{\mathrm{PC}} = O_i \, (t_{ij} / \sum_s t_{is})$
Attraction	(α, γ)	$t_{ij} = {O_i}^{\alpha} \, {d_{ij}}^{-\gamma}$	$T_{ij}^{\mathrm{AC}} = (t_{ij} / \sum_s t_{sj}) \, D_j$
Doubly	γ	$t_{ij} = {d_{ij}}^{-\gamma}$	$T_{ij}^{\mathrm{DC}} = A_i O_i t_{ij} B_j D_j$

Table 3.1: Overview of the constrained models within the gravity family.

parameters $\boldsymbol{\theta}^{\mathrm{AC}} = (\alpha, \gamma)$. Replacing the reduced affinity function into Equation (3.7a) yields

$$p_{ij}^{\mathrm{AC}} = \frac{{O_i}^{\alpha} / {d_{ij}}^{\gamma}}{\sum_s {O_s}^{\alpha} / {d_{sj}}^{\gamma}} \,, \tag{3.17a}$$

$$\langle T_{ij}^{\mathrm{AC}} \rangle = p_{ij}^{\mathrm{AC}} \, D_j \,. \tag{3.17b}$$

Now the empirical flows play the opposite roles compared to the production-constrained model, with the D_j terms ensuring the constraints, and the O_i terms appearing inside the probabilities. There is a nice symmetry to these two models when we use the empirical flows inside the affinity function.

3.2.4 The doubly-constrained model

From Equation (3.9), it is now clear that we can multiply each row and each column of the affinity matrix (t_{ij}) by a constant without affecting the final form. Therefore, when we employ the gravity law, Equation (3.13), the relevant term inside the affinity function in the case of the doubly-constrained model is ${d_{ij}}^{-\gamma}$ and the model is specified with only one parameter. The empirical flows appear only in the constraint terms, completing the nice symmetry we had for singly-constrained models.

A comparison of the different forms of gravity law can be found in Table 3.1.

3.2.5 Calibration of the parameters

We now address the question of how to calibrate the parameters inside the gravity law. We find it helpful to define the set of indices for which the data records a nonzero flow as

$$\Omega = \{(i, j) \in \mathbb{N}^2 : T_{ij} > 0\} \,. \tag{3.18}$$

We start with the unconstrained model, $\langle T_{ij}^{\mathrm{UC}} \rangle = K {O_i}^{\alpha} {D_j}^{\beta} / {d_{ij}}^{\gamma}$. Taking the logarithm of this expression and assuming that it fits the observations, we find

$$\tilde{K} + \alpha \log O_i + \beta \log D_j - \gamma \log d_{ij} = \log T_{ij} \,. \tag{3.19}$$

The right hand side is only defined for $(i, j) \in \Omega$, so we recognise a linear system of $|\Omega|$ equations on the four unknowns $(\tilde{K}, \alpha, \beta, \gamma)$. In any useful scenario, $|\Omega| > 4$ and we have an over-constrained problem. We can therefore find the parameters as the solution to a linear least-

squares problem. This process is followed, for example, by Balcan et al. [7], Masucci et al. [65], Yang et al. [126], and Dong et al. [26].

Writing the production-constrained model in full,

$$\langle T_{ij}^{\mathrm{PC}} \rangle = O_i \frac{{D_j}^{\beta} / {d_{ij}}^{\gamma}}{\sum_k {D_k}^{\beta} / d_{ik}^{\gamma}}, \tag{3.20}$$

we can see that the simple trick of using the logarithm to get a linear system will not work any more; the sum in the denominator makes it impossible to factorise the terms with β or γ. That being said, we can still write the loss function

$$\mathrm{D}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i,j \in \Omega} \left(T_{ij} - \langle T_{ij}^{*} \rangle \right)^2, \tag{3.21}$$

where $\mathbf{T}^{*}$ is the production-constrained model or any other model, and minimise this expression with respect to the model parameters $\boldsymbol{\theta} = (\beta, \gamma)$. This is a nonlinear least-squares problem and there are good routines in most programming languages specifically designed to find its solution. Our code is written in Python, and we make use of the `least_squares` function within the `scipy.optimize` package [122]. By considering only the nonzero entries (when we restrict the sum to the set of indices in Ω), we are biasing our model on purpose to produce larger entries overall. This is not an issue because in the application that we have in the following chapter we only care about comparing the edges that are observed, that is, the entries in $\mathbf{T}$ which are nonzero.

The attraction-constrained model is dealt with in a similar way, minimising the loss function with respect to the parameters $\boldsymbol{\theta} = (\alpha, \gamma)$. Finally, to find the optimal parameter γ in the doubly-constrained case, we use a fixed point scheme in which we apply the nonlinear least-squares procedure, and we solve for the A_i and B_j balancing factors as part of each iteration of the optimisation routine.

To the best of our knowledge, this approach has not been used before. Simini et al. [111] find the model parameters (of a more complicated 9-parameter model for which the short and the long distances are fitted separately) by minimising instead the mean-squared logarithmic error,

$$\mathrm{LD}(\boldsymbol{\theta}) = \frac{1}{|\Omega|} \sum_{i,j \in \Omega} \left(\log T_{ij} - \log \langle T_{ij}^{*} \rangle \right)^2, \tag{3.22}$$

while Lenormand et al. use the CPC (defined below in Section 3.5) as the objective function to maximise [58].

3.3 The radiation model

The radiation model is a parameter-free trip distribution model proposed by Simini et al. [111]. It belongs to the intervening opportunities family, and as such it predicts that the flow between locations does not depend on the absolute distance between them but rather on their spatial arrangement, and hence on the spatial distribution of *opportunities*.

The *opportunities* available at each location measure the relative importance of the different places. For commuters, it makes sense to consider the number of job openings at each location, although this is a piece of information that is generally not easy to obtain. This is why other variables are used as a proxy for *opportunities*: in the original formulation of the radiation model, Simini et al. choose the population at each location; here we select the total number of trips that arrive at each place (an approach that is also followed by Masucci et al. [65] and Piovani et al. [94]).

The IO matrix $\mathbf{S} = (S_{ij})$ stores the count of opportunities that are located closer to the origin than a reference distance. Formally,

$$S_{ij} = \sum_{\substack{k \neq i \\ d_{ik} < d_{ij}}} D_k \,, \tag{3.23}$$

gives the opportunities (measured by the total incoming flow variable) located strictly within a circle centred at the origin i and with radius d_{ij}, excluding the opportunities already present at the origin.

The radiation model is inspired by the physical process of radiation: if a particle (representing a trip) is emitted from i, the probability that it is absorbed at j depends on the particle not being absorbed anywhere else that is closer to i [111]. If the absorbance is measured using the total incoming flows in $\mathbf{D}$, the probability can be written in closed form as

$$p_{ij}^{\infty} = D_i \left[\frac{1}{D_i + S_{ij}} - \frac{1}{D_i + D_j + S_{ij}} \right] . \tag{3.24}$$

This expression was derived in the limit of infinitely many locations. For a finite system, it is easy to verify that $\sum_j p_{ij}^{\infty} = 1 - D_i/M$, where M is the total number of trips, so here we add the finite-size correction term of Masucci et al. [65]

$$p_{ij}^{\mathrm{Rad}} = \left(1 - \frac{D_i}{M}\right)^{-1} \frac{D_i D_j}{(D_i + S_{ij})(D_i + D_j + S_{ij})} \,, \tag{3.25}$$

thus ensuring that the rows of the probability matrix sum up to one. We recognise a production-constrained model, and the expected number of trips can be found using the familiar expression $\langle T_{ij}^{\mathrm{Rad}} \rangle = O_i \, p_{ij}^{\mathrm{Rad}}$.

3.3.1 Other constrained versions

From Equation (3.25), we can work our way back in order to define an affinity function and build the other constrained models. Even if these other models would not have the same interpretation as the production-constrained radiation model, and the assumptions of the particle radiation process would not be met, the encompassing framework of the family of constrained models justifies that we also look at them. We define

$$f_{ij} = \frac{1}{(D_i + S_{ij})(D_i + D_j + S_{ij})} \,, \tag{3.26}$$

Constraint type	Relevant params.	Affinity function	Prediction (exp. value)
Production	-	$t_{ij} = D_j\, f_{ij}$	$T_{ij}^{\mathrm{PC}} = O_i\ (t_{ij}/\sum_s t_{is})$
Attraction	-	$t_{ij} = D_i\, f_{ij}$	$T_{ij}^{\mathrm{AC}} = (t_{ij}/\sum_s t_{sj})\, D_j$
Doubly	-	$t_{ij} = f_{ij}$	$T_{ij}^{\mathrm{DC}} = A_i O_i t_{ij} B_j D_j$

Table 3.2: Overview of the constrained versions of the radiation model (the original formulation is production-constrained).

and we use this quantity to write the basic affinity function of the radiation model as

$$t_{ij} = D_i D_j\, f_{ij}\,. \tag{3.27}$$

We note that unlike the affinity function for the gravity family of models, this function only involves the vector of incoming trips $\mathbf{D}$. This is a consequence of the way an *opportunity* was defined.

The attraction-constrained model is written as $\langle T_{ij}^{\mathrm{AC}}\rangle = p_{ij}^{\mathrm{AC}} D_j$, where $p_{ij}^{\mathrm{AC}} = t_{ij}/\sum_s t_{sj}$. Dropping irrelevant constants, the affinity function in this case is equal to $D_i\, f_{ij}$. Finally, the doubly-constrained affinity function (up to irrelevant constants) is simply f_{ij} and the OD matrix can be obtained in the usual way, $\langle T_{ij}^{\mathrm{DC}}\rangle = A_i O_i\, f_{ij}\, B_j D_j$. Table 3.2 summarises this framework.

3.4 Different notions of distance

Up until now, we have not clarified how we measure distance. Naturally, this is a question that we need to address because in real life we need to specify whether we measure distance in meters, kilometres, etc., and whether the choice affects the modelling in any way.

This question is closely related to whether our model properly deals with dimensions, and it is one of the reasons we started looking at non-dimensional models. If a model is non-dimensional, then the units we use to measure distance will be irrelevant. Thus, for all the constrained models it does not make a difference what we use for our input distance matrix — and for the radiation model, even before we use constraints, we can see that the units of the distance are irrelevant for calculating the intervening opportunities matrix, Equation (3.23).

This opens up the possibility of using not just physical distance but other measures for the "cost of travel", a term that is employed by Wilson [124] and many geographers. For example, we can calculate the driving distance or driving time, and we can even employ an abstract distance in a latent space, like Murray et al. [69] who combine a node2vec embedding [42] and the gravity model to study travel itineraries and the mobility of scientists.

We explored the possibility of querying a web API for the driving time or driving distance between locations,[4] but we found it too impractical for our real-world datasets that have more than a hundred or so nodes because we were limited on the number of queries that we could do for free (and the number needed scales quadratically).

[4]We used the HERE Routing API, a service provided by a consortium of automotive companies, available at https://developer.here.com/products/routing.

A solution that was pointed out to us and that we are keen to develop in the future, is to obtain the road network and calculate the distance on it, assuming that the shortest path is taken and perhaps that the maximum driving speed is allowed. Obviously, this will underestimate the typical driving time because we will not be able to get traffic conditions (something that the web API can account for) but it will still give us a more realistic answer than if we ignore the topography altogether, which is our current best solution.

In the remainder of the chapter, we select the great-circle distance to measure travel cost. If a point $\mathbf{x}$ located on the surface of the Earth has longitude and latitude (λ_x, ϕ_x), measured in radians, the great-circle or Haversine distance separating $\mathbf{x}$ from $\mathbf{y}$ is calculated as

$$D(\mathbf{x}, \mathbf{y}) = 2R \arcsin\left[\sqrt{\sin^2\left(\frac{\phi_x - \phi_y}{2}\right) + \cos(\phi_x)\cos(\phi_y)\sin^2\left(\frac{\lambda_x - \lambda_y}{2}\right)}\right], \tag{3.28}$$

where $R = 6371$ km is the radius of the earth.

For the network of commuting in the UK, we use the centroid of the geographies of local authorities (calculated using the `shapely` package in Python [37]) to represent each node's location.

For the retail network, we did a good amount of work to combine different data sources and obtain the coordinates for 3,152 stores in all four countries of the United Kingdom. Unfortunately, the number of unique coordinates is only 2,563 (81.3% of the total) because large stores that have a petrol station nearby can be given the same coordinates. When we apply the gravity law to the network below, we set the distance between pairs of different stores that have the same coordinates to be 1km. This is arbitrary and by no means an ideal solution, but we did not find a better solution. We experimented using other values than 1km and the results change in subtle ways: the radiation model is unaffected if we choose a value that is smaller than the typical distance to other stores (this changes in different situations) but the gravity model predictions are much more sensitive and if we choose too small a value the predictions grow very large (in the next chapter we mitigate for this because we develop a significance tests which makes the absolute size of the predicted flows less important).

3.5 Goodness-of-fit measures

We now change topics and discuss tools to gauge the fit of the different models to the empirical OD matrices. Perhaps the most commonly used goodness-of-fit measure — at least in the Physics and Applied Mathematics literature — is called the common part of commuters (CPC) [59], based on the Sørensen index for ecology. It is defined as

$$\mathrm{CPC}\left(\mathbf{T}^{(1)}, \mathbf{T}^{(2)}\right) = \frac{\sum_{i,j} \min\left(T_{ij}^{(1)}, T_{ij}^{(2)}\right)}{\left(\sum_{i,j} T_{ij}^{(1)} + T_{ij}^{(2)}\right) \Big/ 2}. \tag{3.29}$$

If the sum of the entries in the two matrices is equal to M (the total number of trips inside an OD matrix), the expression above simplifies to

$$\mathrm{CPC}\left(\mathbf{T}^{(1)}, \mathbf{T}^{(2)}\right) = \frac{1}{M} \sum_{i,j} \min\left(T_{ij}^{(1)}, T_{ij}^{(2)}\right) . \tag{3.30}$$

For fixed i, j, taking the minimum between the entries in each matrix is a way to measure their agreement and the CPC score varies from 0 to 1, respectively when the nonzero entries of the matrices do not match and when the two matrices agree perfectly. We note however, that taking the minimum between pairs of entries does not tell us how far apart they could be.

The root-mean-square error (RMSE) does take into account the distance between corresponding entries. It is significantly less common that the CPC but we do still encounter it from time to time in some mobility studies. It is calculated as

$$\mathrm{RMSE}\left(\mathbf{T}^{(1)}, \mathbf{T}^{(2)}\right) = \frac{1}{N} \sqrt{\sum_{i,j} \left(T_{ij}^{(1)} - T_{ij}^{(2)}\right)^2} . \tag{3.31}$$

We note that this measure is closely related to our loss function, Equation (3.21), with the difference being that for the loss function we consider only the subset of indices Ω for which the observations are nonzero. Sometimes a normalization term $1/M$ is added as well.

3.6 Application to real datasets

We now apply the gravity and the radiation models to two real-world datasets in the UK, the network of commuting — the work-related movement of people recorded in the 2011 census at the level of 404 local authorities — and a network constructed from the trajectories of customers of one of the largest retail chains in the country.

3.6.1 Network of commuting in the UK

We study first at the calibration of the parameters using our proposed loss function, Equation (3.21), and the nonlinear least-squares scheme. We compare this with the parameters found with other objective functions, in particular the log-deviation of Simini et al., Equation (3.22), and the maximum CPC employed by Lenormand et al. in Ref. [58].

Since the singly-constrained models have two free parameters, we can visually inspect what their optimisation landscapes look like. The results are shown in Figures 3.1 and 3.2, respectively for the production-constrained and attraction-constrained models. The landscape for the loss function $\mathrm{D}(\boldsymbol{\theta})$ that we propose can be seen in each figure in Panel A.

For the production-constrained model, we find that the optimisation landscape is locally concave for our loss function and locally convex for the CPC (in Panel C), with the valley of the first and the mountain of the latter approximately mirroring each other. Our method locates the minimiser of $\mathrm{D}(\boldsymbol{\theta})$ (indicated using a white cross) at $\gamma = 2.23$ and $\beta = 0.91$. The CPC maximisation (black square) gives a comparable value for β, but a significantly larger γ.

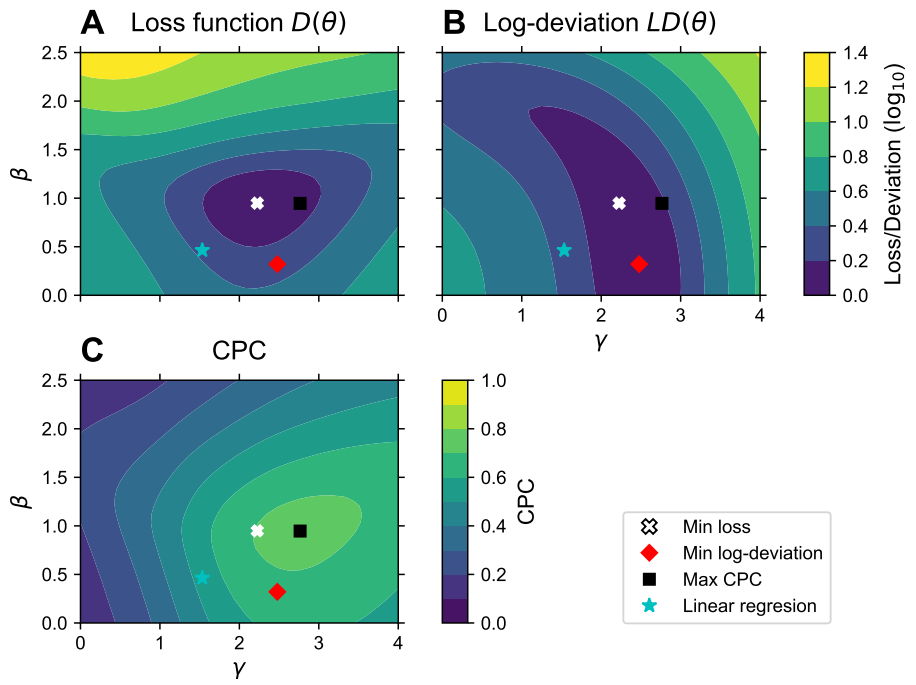


Figure 3.1: Optimisation landscapes for the production-constrained gravity fit to the commuting network. The points correspond to the optimal values of the (γ, β) parameters according to the different methods and their position is the same across the different panels.

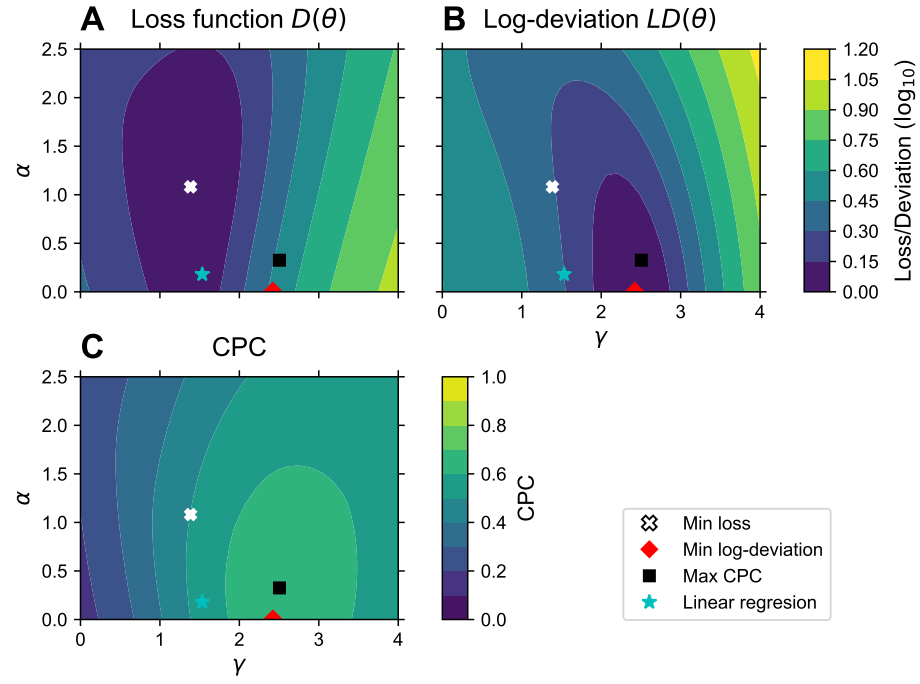


Figure 3.2: Optimisation landscapes for the attraction-constrained gravity fit to the commuting network. The points correspond to the optimal values of the (γ, α) parameters according to the different methods.

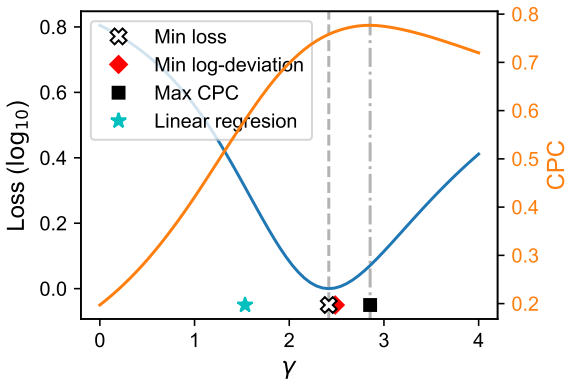


Figure 3.3: Loss and CPC landscapes (in 1-D) for the doubly-constrained gravity model. The value of the different parameters is indicated with the symbols at the bottom, and the vertical lines indicate the minimiser for the loss function D($\boldsymbol{\theta}$) (white cross) and the CPC maximiser (black square).

Panel B shows the landscape for the log-deviation LD($\boldsymbol{\theta}$) and we notice that the shape is more degenerate, with a relatively flat portion which encompasses both its minimiser (red diamond) and the minimiser of D($\boldsymbol{\theta}$). For completeness, we also show the values that would be estimated using a simple linear regression, but we have already established that those parameters are not appropriate for a constrained model. Still, some authors still use these set of parameters even when their model is implicitly constrained.

For the attraction-constrained case, Figure 3.2, we observe a flatter landscape in Panel A that could indicate some amount of degeneracy. Still, the optimisation procedure locates the overall minimiser at $\gamma = 1.47$, $\alpha = 0.95$. The log-deviation, shown in Panel B, has a minimiser well within the negative axis and we have to constrain the minimisation routine so that we ensure $\alpha \geq 0$. This then leads us to find a constrained minimiser (red diamond) that has exactly $\alpha = 0$, which in practice means that the gravity prediction does not depend on the O_i values and only depends on the distance. This probably tells us that the combination of the attraction-constrained gravity model and the log-deviation are not appropriate to model this dataset. For its part, the CPC landscape is also tilted towards the bottom of the image, and the maximiser has a significantly smaller α than our method finds. Whether this is a bad thing, we will decide using a cross-validation analysis further below.

The results for the 1-D doubly constrained model can be seen in Figure 3.3. For clarity, we only compare our loss function and the CPC (the log-deviation curve is very similar to the former). Our method and the log-deviation of Simini et al. find values of gamma just next to each other — our method finds $\gamma = 2.42$. As we had before, the CPC maximiser is located to the right of this.

We would like to compare the performance of the gravity models with the different combinations of parameters that we found. However, we cannot use the goodness-of-fit metrics directly. The reason for this is that we already employed them to find the parameters. In other words, we already know that the CPC maximiser will maximise the CPC, and it is very likely that the

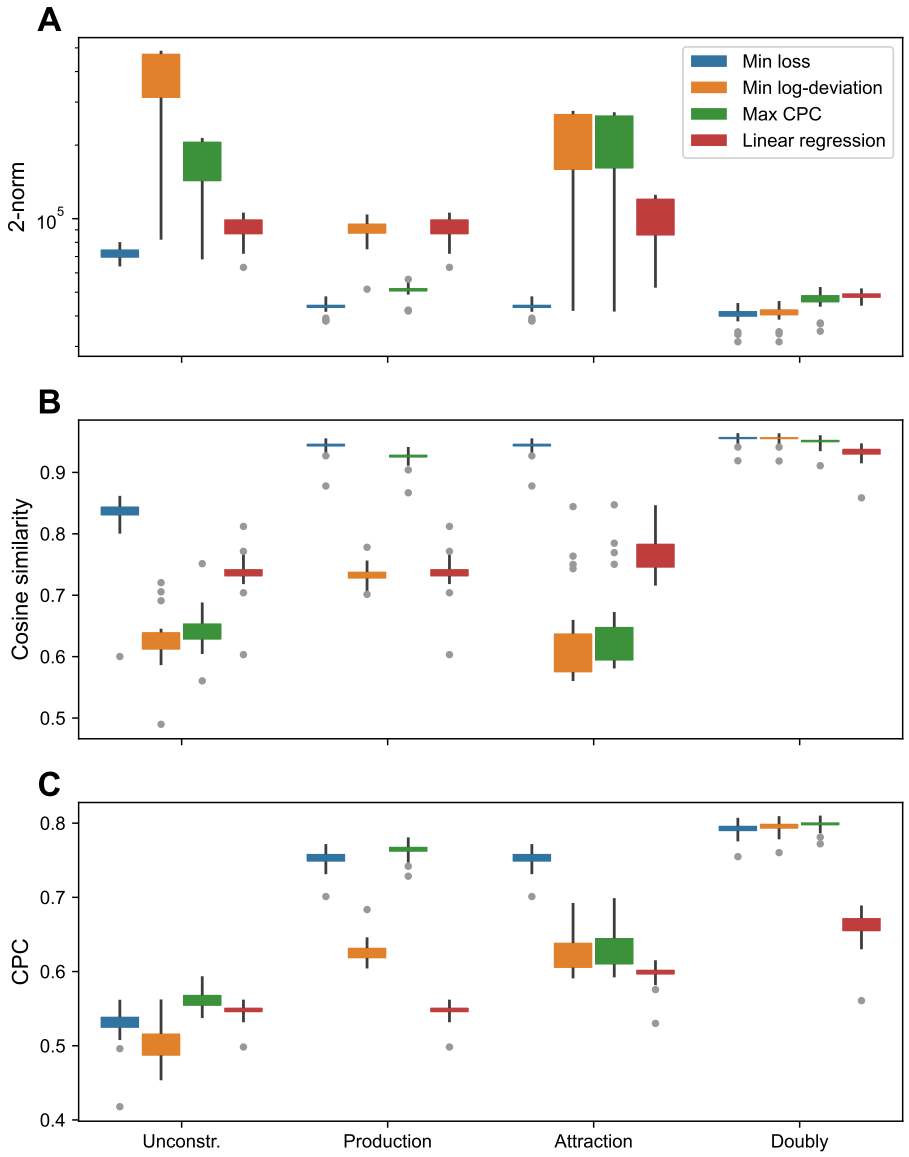


Figure 3.4: Cross-validation comparison of the different parameter fitting procedures. We compare from left to right the different constrained models. The training data is about 80% of the data and the testing data is around 20%. The actual numbers vary because what we fix is the number of nodes selected to test the data. We repeat each experiment 20 times.

		UC	PC	AC	DC
Gravity (min loss function)	CPC	0.435	0.706	0.501	**0.758**
	RMSE	521	298	459	**249**
Radiation	CPC	0.516	0.575	0.601	0.659
	RMSE	702	717	682	497

Table 3.3: Comparison of the goodness-of-fit of the different models in terms of the common part of commuters (CPC) and the root-mean-square error (RMSE) for the network of commuting in the UK. The results for the best model are in bold.

parameters $\boldsymbol{\theta}$ that we found using our method are close to the minimiser of the RMSE (since the difference is only the support Ω).

To do a fair comparison, we perform a cross-validation in which we use part of the data to fit the parameters and we test the performance on the remaining data. We select n nodes at random and we reorder the OD matrix so that we place them behind the other nodes. We use the subset of the matrix corresponding to the first $N - n$ nodes to fit the model parameters, and then we benchmark the model's prediction for the entries that correspond to the set of n selected nodes. For the network of commuting, we set $n = 40$ and with this choice we split roughly 80-20% percent of nonzero entries respectively into the training and testing sets. We repeat the cross-validation experiment 20 times.

The results can be seen in Figure 3.4. We compute the 2-norm of the difference between the data and the model predictions, and also the cosine similarity. On average, our method results in the lowest 2-norm difference and the highest similarity for all types of constrained models. Though there are particular instances when a different parameter calibration method can outperform the minimum loss, our method is most consistently the best (or close to the best) method.

Furthermore, we can now establish whether it is in fact a bad thing that the α parameter be so low for the case of the attraction-constrained model: the minimum log-deviation that sets $\alpha = 0$ and the maximum CPC that results in $\alpha = 0.33$ lead to some of the worst results overall, with a particularly high interquartile range for the 2-norm and low cosine similarity.

Now that we have settled that the best way to find the parameters of the gravity model is by minimising the loss function, Equation (3.21), we compare the gravity and the radiation models and each of their constrained formulations.

The results for the common part of commuters and the root-mean-square error can be found inside Table 3.3. Focusing on each row, we can see that the doubly-constrained models outperform both of the singly-constrained ones, and in turn these outperform the unconstrained model. We also obtain better results overall with the gravity model than with the radiation model and this holds for each of the constrained formulations (the columns in the table).

When we compare the doubly constrained models, we notice that a small difference in the CPC can still translate into a very large difference in the RMSE: indeed, the CPC score for the DC gravity model is 15% large than for the DC radiation model, but the RMSE score is essentially half. We said before that the CPC does not quantify the entry-wise spread of the

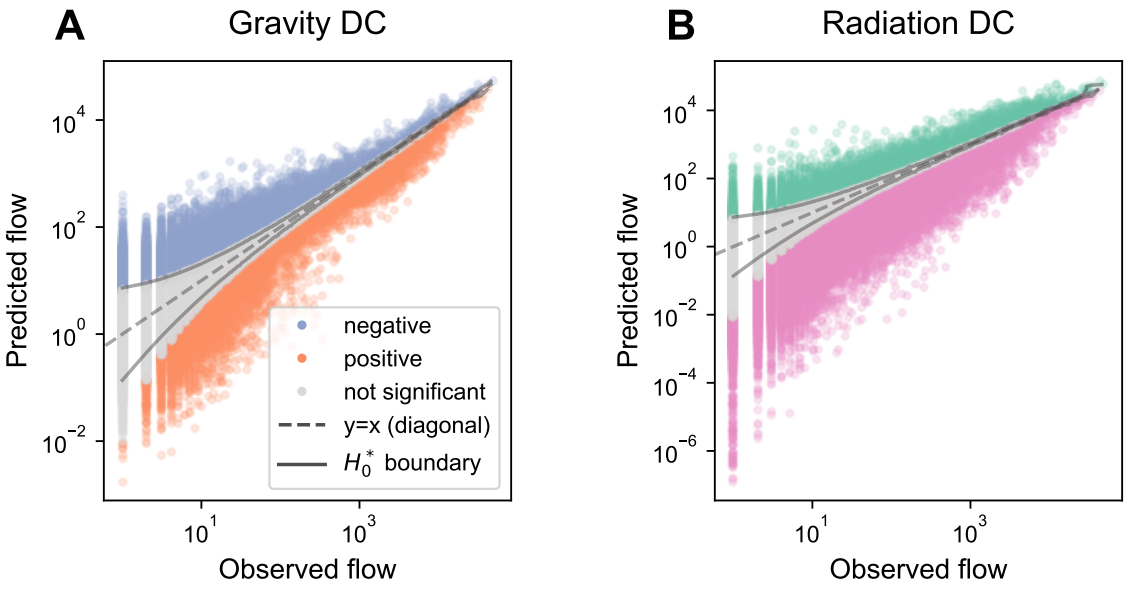


Figure 3.5: Doubly-constrained gravity and radiation models applied to the commuting network in the UK. The colours indicate the sign of the edges as determined with the spatial backbone that will be introduced in the next chapter.

models being compared, and we are likely seeing a result of this. We should therefore interpret the CPC with caution; attractive as it might be to have a measure that varies between 0 and 1, it is often assumed that a numerical quantity scales linearly and this is simply not the case for the common part of commuters.

As is standard in the literature, in Figure 3.5 we visually inspect the fit of our models using a scatter plot to compare the observation T_{ij} and the expected value $\langle T_{ij}^* \rangle$ for all pairs of locations that have nonzero data — the colours indicate the sign of the edges according to the spatial backbone that will be introduced in the following chapter. Note that the axes are scaled logarithmically. The dashed line is the diagonal $y = x$ and our predictions can be off by several orders of magnitude, a feature that is standard for all types of mobility models applied in many different settings (the mobility models leave a large variance unaccounted for). This is most prominent on the left side of the plots and it affects less the right side of the plot because the largest flows play a more important role when the models are calibrated.

3.6.2 Retail network

We study the anonymised dataset of shopping records generated by over 1.4 million customers during fiscal year 2017.

In the OD matrix built from the census data, every trip represents a single person. We approach the retail dataset in a different way. For each customer, we take the sequence of visited stores and we count successive visits as trips between locations (over the period of interest). For example, if a given customer visited three stores in the following order (A, B, B, C), this will contribute one trip from store A to store B, one self-visit from B to B (that is not kept in the OD matrix) and one trip from B to C. To build our aggregated OD matrix, we sum the counts

		UC	PC	AC	DC
Gravity (min loss function)	CPC	0.257	0.386	0.376	**0.726**
	RMSE	296	249	249	136
Radiation	CPC	0.397	0.674	0.551	0.706
	RMSE	337	150	265	**133**

Table 3.4: Comparison of the goodness-of-fit of the different models in terms of the common part of commuters (CPC) and the root-mean-square error (RMSE) for the retail network. The results corresponding to the best performing model are in bold.

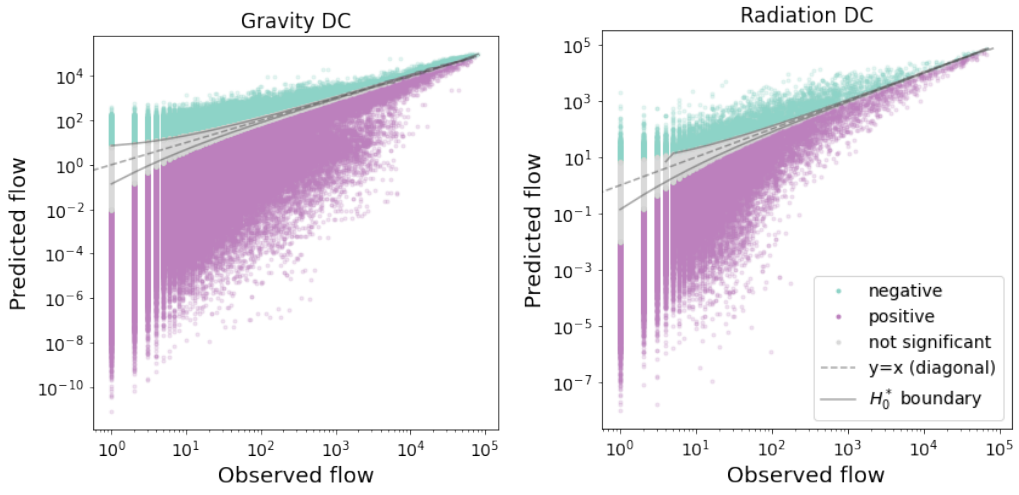
for all the customers over all the stores.

Obviously, this procedure has its limitations. Perhaps the most important one is that we do not account for the different time intervals between the visits to the retail chain. We can suppose that only a day elapsed between the visits to the stores A and B in the example above, while the customer visited shop C two weeks after the second visit to shop B. If this were true, the trip $A \to B$ could signal a stronger interaction between the two shops than $B \to C$. We could think of adding a discount factor that decreases with the time delta, but this has the disadvantage that we need to calibrate the representative time (probably for each customer) and we leave this for future work.

We could think that the shopping flows in this way constructed look nothing like the spatial interactions normally described with the mobility models, but it appears that the gravity and the radiation model fit the shopping OD matrix at least as well as the commuting data, as attested by the goodness-of-fit results presented in Table 3.4: the CPC of both the gravity and the radiation doubly-constrained modes are above 0.7. In this way, the fact that trips can no longer be attributed to a single customer might be exactly what we need to put the data in the right format. Regarding the issues we have noted with the nature of the CPC, we find here that the singly-constrained radiation models score differently according to the CPC, but they give the same RMSE (rounded to the nearest integer).

For the gravity model, we get a significantly better fit when we use the doubly-constrained model compared to the singly-constrained ones but, interestingly, the improvement is much more modest for the radiation model because the production-constrained model already fits the data very well. We pay a much lower computational cost for calculating the production-constrained model because we do not need to find the balancing factors, so this remark could be interesting for prototyping purposes. Furthermore, comparing these results with the ones we obtained for the network of commuting, we see almost the opposite behaviour in that there it was the production-constrained gravity model that performed almost as well as its doubly-constrained counterpart.

It is interesting to take Tables 3.3 and 3.4 together: while from the former we conclude that the gravity model clearly outperforms the radiation model, in the latter the performance is similar for the two models (the CPC is more favourable for the gravity DC model but the RMSE gives almost identical results). This is interesting in light of the fact that the two datasets cover the same geographical area, therefore this speaks about the different nature of the mobility processes being described.



(a) (b)

Figure 3.6: Doubly-constrained models applied to the retail OD matrix: (a) gravity model (b) radiation model

We hypothesise that in the situation of deciding where to shop for groceries, the "mechanistic" radiation model — a type of intervening opportunities model — could be comparatively better suited than in other settings. Indeed, people often shop in the store that is closest to them and they will only travel further to do a bigger shop or to find a specific product. This sounds like the ideal situation that the framework of intervening opportunities aims to model.

3.7 Final considerations

By using Wilson's family of constrained models, we have turned the gravity law into three separate models that have a reduced number of parameters. But what we lost in expressiveness is compensated by the additional constraints, as attested by the increasing goodness-of-fit the more constraints we use.

Constrained models are commonly employed (the production-constrained model is by far the most common), and as such the use of the empirical flows are also widespread. However, practitioners still seem reluctant to use the flows to represent the propulsiveness and the attractiveness of the locations, and many uses of the gravity law still rely on the population, a country's domestic product, etc. These variables, although correlated with the origin-destination data, do not come from the data itself and as such they are somewhat arbitrary. But more importantly, they add more inputs than required since the empirical flows are also used in order to constrain the models. What we have shown is that, whenever it is practical, the empirical flows should be preferred as the canonical propulsiveness and attractiveness variables because this is a more efficient use of the data, and it also respects the symmetry of the constrained formulations.

We must recognise, however, that constrained models are not always practical because they assume that we have closed systems, i.e. systems that have zero incoming and zero outgoing flows from locations that are not part of the system (think the Republic of Ireland or continental Europe for the networks that we have described before). Relaxing the two requirements turns the

doubly-constrained model (describing a fully closed system) into the production or attraction-constrained models (respectively, without flows originating outside the system or incoming from outside) and then a fully unconstrained model which can be employed even when the system is completely open. Open or partially open systems are the only way forward in some scenarios. For example, attraction-constrained models will be selected to estimate the demand of new stores given their size or their availability of products (the attractiveness values) but without being able to measure the number of customers.

Here we make the assumption that the systems that we study are, too a good approximation, closed. We saw that the doubly-constrained model leads to the best fit overall so it would seem that they should always be preferred to singly-constrained models in the case of the gravity law. However, for the commuting network we found that the production-constrained model gives competitive results with the CPC scoring above 0.7 and the RMSE below 300 (unlike the attraction-constrained model which is closer to the unconstrained formulation). We might then find ourselves in situations in which the cost of the lower accuracy achieved with the production-constrained model is worth paying due to its significantly lower computational cost compared to the doubly-constrained model.

Finally, we mentioned that there is a misconception that doubly-constrained models are deterministic, but we showed that it is still possible to define probabilities and use them inside a statistical model. All in all, we hope to have clarified some of the questions that arise when using a model as popular and useful, though still confusing, as the gravity law.

Chapter 4

Stochastic block modelling of spatial networks with signed backbones

The commuting and the retail network are examples of spatial networks. This means that normal methods for community detection are not particularly well suited to discover meaningful groups of nodes in them. For example, the modularity quality function with the Newman–Girvan null model assumes that any two nodes can be connected with a probability that depends solely on their degree, and this assumption is unlikely to hold for many spatial networks. Both for modularity and for other flow-based methods, the application of simple methods will yield communities of nodes that tend to be geographically clustered [27, 36, 43]

Methods to find space-corrected communities have been proposed [14, 27, 106], but so far they have mostly focused on finding a more suitable null model to be used inside the modularity function (we review these proposals in the following section). However, a spatial version of the stochastic block model is still something that eludes us.

In this chapter, we propose a novel way to find space-corrected communities using an SBM. Our method relies on a pre-processing step in which we build a network where spatial effects have been subtracted. We call this the spatial backbone of our network. The name is inspired by the backbone extraction problem in which the goal is to find the most important edges in a weighted network. This is usually performed in order to convert a dense weighted graph into a sparse binary graph, which simplifies the analysis. Sparsifying the network, however, is a side effect of computing the spatial backbone, but the main motivation for us is constructing a new network where the most important edges — the edges that are kept — are those that we cannot explain using a model for the spatial interactions. We can then leverage the spatial backbone and employ the stochastic block model to find a space-corrected partition of the nodes.

4.1 Modularity-based methods for space-correction

Let us suppose for the time being that the empirical OD matrix $\mathbf{T}$ is symmetric and let $M = \sum_{i,j} T_{ij}/2$ denote the total edge weights. In this context, we have $O_i = D_i$ and we simply denote the in-strength equal to the out-strength by k_i so that we can use a familiar notation.

The null model P_{ij} in the modularity function should reflect what we believe is a good model for our network. As discussed in Section 2.2.2, the Newman–Girvan null sets $P_{ij}^{\mathrm{NG}} = k_i k_j / 2M$ so that probability of connection between nodes depend solely on their degree, but this is unlikely to hold if the network is influenced by space. For this reason, there has been different proposals of spatial null models.

Expert et al. were the first to propose a null model adapted to spatial networks [27]. For nodes i and j a distance d_{ij} apart, their model is

$$P_{ij}^{\mathrm{Spa}} = n_i n_j f(d_{ij})\,, \tag{4.1}$$

where n_i is a measure of the importance of location i (for example, its population). The authors estimate the deterrence function f numerically from the data using a binning on the distance, but we already saw that for the gravity law we more commonly find f to be a monotically decreasing function of its arguments such as $f(d) = d^{-\gamma}$ or $f(d) = \exp(-d/\ell)$.

An alternative null model by Liu et al. [61] sets

$$\tilde{P}_{ij} = \frac{k_i k_j f(d_{ij})}{\sum_s k_s f(d_{is})}\,, \tag{4.2}$$

$$P_{ij}^{\mathrm{Dist}} = (\tilde{P}_{ij} + \tilde{P}_{ji})/2\,. \tag{4.3}$$

The expression for $\tilde{P}_{ij}$ corresponds to the production-constrained formulation of the gravity model and the null model is the average between the outgoing and incoming flows. Liu et al. explore two different choices for the deterrence function: $f(d) = \exp[-(d/\sigma)^2]$ and $f(d) = [1 + (d/\sigma)^2]^{-1}$.

Lastly, Sarzynska et al. propose an altogether different form that is based on the radiation model [106]. If we denote by S_{ij} the sum of population of all the locations strictly within the annulus centred at i and defined by $0 < r < d_{ij}$, then the flow predicted by the radiation model is

$$\tilde{T}_{ij} = k_i \frac{n_i n_j}{(n_i + S_{ij})(n_i + n_j + S_{ij})}\,, \tag{4.4}$$

$$T_{ij}^{*} = (\tilde{T}_{ij} + \tilde{T}_{ji})/2 \tag{4.5}$$

The null model P_{ij}^{Rad} is obtained from the predicted flows using the same type of binning that Expert et al. used to estimate the deterrence function (we do not include the expression here but refer the reader to reference [106]).

4.1.1 Limitations of spatial modularities

It is more or less obvious from the forms given above that the symmetrisation of the trip-distribution models is an inconvenience, and it is only employed so that the standard modularity maximisation framework can work with the spatial networks that had to be made undirected themselves. Modularity can be extended to directed network [3], but it is interesting that none of the authors above tried to make their method directed. From a modelling point of view, we

believe that this is a gross omission because both the trip-distribution models and the spatial networks are normally directed and for good reason: the number of commuters that go into the City of London from Brighton is much larger than in the opposite direction.

As explained in Section 2.2.2, modularity maximisation will tend to overfit a network and will find structure even tree-like networks which intuitively are not organised into communities. But perhaps the most serious limitation of modularity-based methods is that they are, by definition, limited to find groups of nodes that are assortative, i.e. densely connected within groups. However, once the natural clustering that occurs due to spatial proximity has been discounted, there is nothing particular about spatial data that justifies that the networks should be organised in an assortative fashion. Going back to our previous example, in terms of groups of nodes, our method should probably divide into different groups the sink nodes with a net inflow (similar to the City of London) and the source nodes with a net outflow (similar to Brighton). This pattern is not assortative and modularity would not be able to find it. The stochastic block models are suited to both directed and undirected networks, so and if we can adapt them to spatial networks they will be ideal to avoid the need to make a trip distribution model symmetric.

Finally, we can also consider the definition of "surprise" that is measured with the modularity function. For a given pair of nodes, surprise will normally be a positive term $A_{ij} - P_{ij}$, which the maximisation is also striving to make as large as possible. But this begs the question, what happens when the term $A_{ij} - P_{ij}$ is negative? Should this not be considered surprising in its own right? Let us now discuss an alternative way of measuring surprise with the backbone of a network.

4.2 The backbone extraction problem for a spatial network

In the backbone extraction problem, the goal is to find the most important edges in a weighted network. Important edges are statistically significant when compared against a null model [110], that is, a model of how the weighted network came to be. When we study a network with spatial interactions, we can adopt any of the mobility models that we studied previously — the gravity and the radiation model — as a model for our network and use it as a point of comparison.

The weights in our spatial network will be taken to be the flows recorded in the empirical OD matrix $\mathbf{T}$. The flows can vary over several orders of magnitude for different pairs of locations, and this raises the question of how to deal with the multiscale organization of the network in a consistent way. Multiscale systems are systems that tend to show different characteristics or behaviours when studied at different scales. We can find multiscale networks even if the distribution of weights does not vary a great deal, but it is more or less clear that if the distribution of weights varies over a large range this will result in a multiscale network.

In a weighted network, a global threshold is sometimes used to define importance and to keep the weights above a suitably chosen magnitude. However, this method penalises the smallest nodes that have a small strength and it also introduces a characteristic scale below which all the information is discarded [110].

A better approach is to compare the weights in the networks against what can be obtained

under a suitable null model, and test for statistical significance. Serrano et al. [110] compare the normalised node weights at a node of degree k (the weight divided by the node's strength) against an assignment of $k-1$ numbers uniformly at random in the unit interval. Neal [70] considers the weights obtained when a bipartite networks is projected, and his null model corresponds to a counting argument of the number of ways in which a value can be obtained. Significant edges can be found for all sizes of nodes (measured, for example, using their strength) and the backbones preserve the multiscale nature of the data.

Here we assume that either of our preferred trip-distribution model is a suitable mechanism for the formation of our spatial network. The entry T_{ij} is compared against our prediction for the expected value, $\langle T_{ij}^* \rangle$, or more generally against the random variable T_{ij}^*. Then, by employing one-tailed significance tests, we determine the statistical significance of the observed weights in order to find those that can be considered smaller or larger than expected according to our model.

4.2.1 Significance test for the observations

The doubly-constrained formulation of either one of our mobility models can be written in either a production- or attraction- constrained form simply by factoring out, respectively, the O_i or the D_j term (see Section 3.2.4). For the pair of locations i and j, let us write $X = O_i$ if we place ourselves in a production-constrained setting and $X = D_j$ if we take the attraction-constrained view instead.

Both the gravity and the radiation model give us a vector of probabilities (outgoing or incoming), and the distribution of the flows involving locations i and j follows a multinomial distribution (either outgoing from i or incoming into j). The univariate marginal T_{ij}^* is binomially distributed and its probability mass function is given by

$$\mathbb{P}(T_{ij}^* = t) = \binom{X}{t} {p_{ij}}^t (1 - p_{ij})^{X-t}\,, \tag{4.6}$$

for integer $0 \leq t \leq X$. The expected value of our random variable is $\langle T_{ij}^* \rangle = X p_{ij}$ and its variance is $\sigma_{ij}^2 = X\, p_{ij}\, (1 - p_{ij})$, where $X = O_i$ or $X = D_j$ as appropriate.

4.2.2 Exact computation

We want to address the question of which entries in the matrix or alternatively which edges in the network are statistically significant. We do so by employing two statistical tests. The null hypothesis that we consider is

$$H_0: \text{ The observed flow } T_{ij} \text{ is binomially distributed with parameters } X \text{ and } p_{ij}. \tag{4.7}$$

Our first test is a one-sided right-tailed test to assess whether the data point T_{ij} stands out

because it is particularly large. The corresponding p-value is calculated as

$$\mathbb{P}\left(T_{ij}^* \geq T_{ij} \mid H_0\right) = \sum_{t=T_{ij}}^{X} \mathbb{P}(T_{ij}^* = t) \tag{4.8}$$

$$= 1 - \sum_{t=0}^{T_{ij}-1} \mathbb{P}(T_{ij}^* = t)\,. \tag{4.9}$$

The observation T_{ij} is deemed significantly large if the p-value is smaller than a predetermined significance level $\alpha = 0.01$. We call the associated edge in the flow network a positive edge (this name is chosen to imply that the difference between the observation and the model prediction is positive).

The second test is a one-sided left-tailed test to determine the significance of particularly small entries. The associated p-value is

$$\mathbb{P}\left(T_{ij}^* \leq T_{ij} \mid H_0\right) = \sum_{t=0}^{T_{ij}} \mathbb{P}(T_{ij} = t)\,. \tag{4.10}$$

We use the significance level from before, and T_{ij} is deemed significantly small if the p-value is smaller than α. We call the associated edge negative.

We note that the two tests are mutually exclusive and an entry will be positive, negative or neither positive nor negative. This third case corresponds to the entries that are deemed statistically not-significant because the null hypothesis cannot be rejected for them. In other words, these are the edges that can be explained by the predictions of our mobility model, and we call these zero edges.

From the statistical significance of the nonzero entries in the empirical OD matrix $\mathbf{T}$, we build a new $N \times N$ matrix $\mathbf{A} = (A_{ij})$ whose entries are defined as

$$A_{ij} = \begin{cases} 1 & \text{if } T_{ij} \text{ is a positive edge,} \\ -1 & \text{if } T_{ij} \text{ is a negative edge,} \\ 0 & \text{otherwise.} \end{cases} \tag{4.11}$$

This signed matrix construction based on positive or negative significance is also employed by Neal [70] and Gursoy and Badur [46]. We find it useful to decompose our matrix as $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$, where both $\mathbf{A}^+$ and $\mathbf{A}^-$ have entries in $\{0, 1\}$ and can thus be viewed as adjacency matrices. We do not consider the zero edges any further because we assume that the mobility model explains the mechanism of how those edges came to be.

4.2.3 Approximate computation

The calculation of the p-values, Equations (4.9) and (4.10), can be computationally expensive and it can be useful to devise an approximate test.

If the total flow X is sufficiently large and p_{ij} is not close to 0 or 1, the binomial distribution

$\mathrm{Bin}(X, p_{ij})$ can be approximated by a normal distribution and we can use a Z-test.[1] We compute the Z-score for entry T_{ij},

$$Z_{ij} = \frac{T_{ij} - \langle T^*_{ij} \rangle}{\sigma_{ij}} \,. \tag{4.12}$$

Our null hypothesis is

$$H_0^* : \text{ The statistic } Z_{ij} \text{ is normally distributed with mean 0 and standard deviation 1.} \tag{4.13}$$

If we let Y be a normally distributed random variable, $Y \sim \mathcal{N}(0,1)$, the p-value of the right-tailed test from before becomes

$$\mathbb{P}(Y \geq Z_{ij} \mid H_0^*) = \Phi(-Z_{ij}) \,, \tag{4.14}$$

where Φ is the cumulative distribution function of $\mathcal{N}(0,1)$, while the p-value of the left-tailed test is

$$\mathbb{P}(Y \leq Z_{ij} \mid H_0^*) = \Phi(Z_{ij}) \,. \tag{4.15}$$

4.3 Stochastic block modelling of spatial backbones

Having found the spatial backbone of a network, we are left with a directed unweighted graph with signed edges, $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$, to which we would like to apply the generative SBM family of models.

The microcanonical SBM formulation introduced in Section 2.2.3 is well suited for this. Indeed, the unified theory on which it is built provides a clear advantage over other SBMs that were formulated with specific use cases in mind [49, 121]. Furthermore, the MDL principle is applicable because we are satisfied with studying point estimates of the posterior distribution (we mitigate for the degenerate optimisation landscape by running the minimisation algorithm several times using different random seeds), and the inference can be done with the *graph-tool* library in Python [91].

We can follow two different paths: either we consider the signed components separately and fit an SBM to $\mathbf{A}^+$ and then $\mathbf{A}^-$ — we end up with two partition vectors, respectively $\mathbf{b}^+$ and $\mathbf{b}^-$ — or else, we view the backbone as a layered network with $C = 2$ layers and employ the independent layers version of the SBM to get a single vector $\mathbf{b}$.

The first alternative might seem like an odd way of proceeding but our reasoning is that the interactions encoded in $\mathbf{A}^+$ and $\mathbf{A}^-$ are meant to capture opposite signals, and it might be reasonable to expect that two independent SBMs will pick them up in a way that agrees. Indeed, we expect that a region of the network with many positive edges will tend to have few negative ones and vice-versa, and therefore we might end up, for example, with the nodes being grouped into assortative blocks inside the positive layer and disassortative ones in the negative layer. Luckily, the stochastic block model is well suited to pick-up both the abundance and lack

[1] As we will see, there will be many flows for which these assumptions will not hold because the probability p_{ij} will be close to zero. We discuss the implications of this further ahead.

of edges between groups.

Irrespective of the path that we follow, beside the vector(s) of block memberships, we also get the two matrices of edge counts $\mathbf{e}^+$ and $\mathbf{e}^-$ which will be crucial to make sense of the network partitions. There will be a difference, however, in that the layered SBM will result in two matrices that are directly comparable because they were obtained for the same groups of nodes while the independent SBMs will not be easily comparable.

4.4 Synthetic benchmarks for spatial community detection

Before analysing real-world datasets, we first employ the space-corrected SBM methodology on controlled experiments with synthetic networks so that we can benchmark the recovery results.

4.4.1 A unified framework

In the community detection literature, there are two types of synthetic networks that are used for testing: the Uniform model of Expert et al. [27] and the Correlated Group Membership (CGM) model [2] of Cerina et al. [14]. The two models are employed with pretty minor modifications to test all the spatial null models proposed for modularity that we reviewed in Section 4.1.

We now discuss these models in a unified manner, separating the two main aspects of a synthetic network model: the spatial arrangement of the nodes and the weight associated to each edge. In doing so, we introduce our own proposal for a more realistic spatial arrangement.

Spatial arrangement of the synthetic nodes

A synthetic model aims to generate networks that have certain properties fixed and other aspects randomised and different for each draw. As mentioned above, the first thing to specify is the placement of the nodes.

Expert et al. [27] draw N points uniformly at random inside a square of length ℓ (Figure 4.1A). The first $n = \lceil N/2 \rceil$ nodes are assigned to the first block and the remaining ones are assigned to the second group. This results in each group being uniformly distributed inside the square.

Cerina et al. [14] study the possibility of a controlled correlation between space and the latent group membership. In their model, there are two centres with coordinates $(\pm\ell, 0)$; around each of them we place $N/2$ nodes (suppose that N is even) at an exponentially decaying distance that has scale parameter ℓ, and at an angle selected uniformly at random. Each node is given an attribute[3] $s_i = \pm 1$ which is correlated with space in the following way: node i is assigned the attribute $\text{sign}(x_i)$ with probability $1 - \epsilon$ or the opposite sign with probability ϵ. Thus, $\epsilon = 0$ perfectly divides the plane in two with $s_i = -1$ *iff* $x_i \leq 0$ ($s_i = 1$ *iff* $x_i > 0$) — this is the case of perfect correlation between space and group membership — and $\epsilon = 0.5$ is a uniformly random assignment in which space is not involved. An example realisation with $\epsilon = 0.15$ can be found in Figure 4.1, Panel C.

[2] The name is ours; the authors do not name their model.

[3] From the node attributes, we obtain the block assignment that has the standard labelling using $b_i = (3+s_i)/2$.

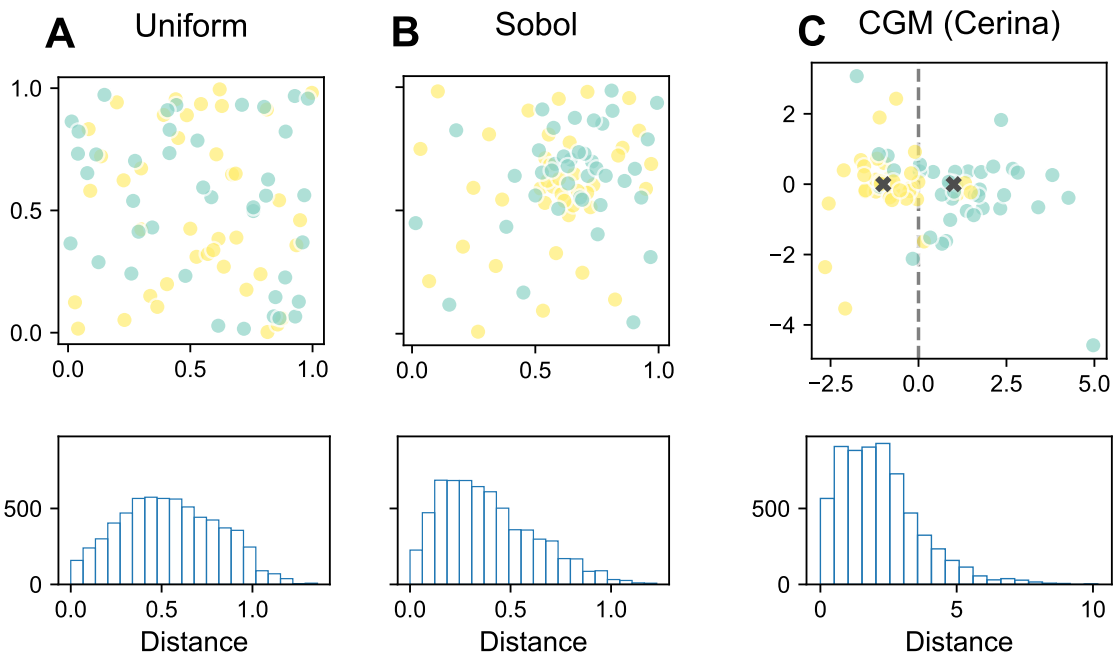


Figure 4.1: Comparison of three different synthetic arrangements. Each column represents a different layout with the generated points in the top row and the resulting distance distribution in the bottom. Each panel has $N = 80$ nodes (manually set for Uniform and CGM) and in each case the characteristic length is set to $\ell = 1$. For the Uniform (Panel **A.**) and Sobol (**B.**) layouts, ℓ is the side of the square; for the Correlated Group Membership model (**C.**), the characteristic length gives the position of the two centres (crosses) and the scale of the exponential distribution. We note that Panel C has a different scale compared to the other two. Furthermore, the correlation parameter is set to $\epsilon = 0.15$ and we can see that a majority of the points are coloured according to the side of the plane that they are in (dashed line) yet $\epsilon > 0$ so there are a few exceptions.

If the attribute s_i is strongly correlated with space, the authors explain in Ref. [14] that normal community detection methods will work in both regimes and, importantly, removing space can lead to incorrect results. However, when the attribute is uncorrelated with space, space-corrected methods are needed if space is dominant. Therefore, the Correlated Group Membership model shows that space can play different roles in community detection, and that space-corrected methods might give misleading results in some situations.

The two spatial arrangements introduced so far have the disadvantage that nodes can be located at an arbitrarily small distance of each other, which is not a behaviour that we observe in real-world systems. We propose a third alternative for the placement of nodes based on Sobol sequences [114] which are low-discrepancy quasi-random numbers that were developed to efficiently calculate Monte Carlo integrals.

We begin, similarly to Expert, with a square of length ℓ. For a parameter p of our choosing, a Sobol sequence gives us a coverage of the square with $n = 2^p$ points (always a power of two) that is quasi-random and prevents two points from getting arbitrarily close. We can then repeat the drawing of points R times to model the multiscale densities of real-world networks that are denser the closer we get to urban centres: after the first n points, we divide the square into four

and discard $n/4$ points (say, those in the upper-right quadrant). Then, we repeat the procedure to fill the empty quadrant of length $\ell/2$ and with $n = 2^p$ points, discarding a quarter of them (say, those in the lower left quadrant if $R > 2$ so as to centre the points with higher densities). We proceed iteratively, drawing and discarding points, until we reach the final iteration and we keep all n new pairs of coordinates. With this procedure, the density increases each time by a factor of 4, and the total number of nodes that we place is $N = n + 3n(R-1)/4$. When we have finished placing all the nodes, we randomly permute them and assign $N/2$ of them to each group. A random realisation with $R = 3$ can be seen in Figure 4.1B. Since we are usually filling four quadrants, we take the convention that the number of nodes that we place is $n = 2^{p+2}$ and this ensures that N is always even ($p = 3$ in the figure above).

In the bottom panels of Figure 4.1, we compare the distribution of distance across the three layouts with the same scale factor $\ell = 1$. We observe that the CGM model has a long tail while the other two layouts do not because the distances are bounded above by $\ell\sqrt{2}$. To get a better sense of how the distributions compare, we generate one hundred networks with each type of layout. The average minimum distance between pairs of points is 0.009 ± 0.005 for the Uniform model, 0.014 ± 0.009 for the Cerina layout and 0.015 ± 0.005 for the Sobol layout; the average mean distances are 0.515 ± 0.018, 1.996 ± 0.138 and 0.391 ± 0.001 respectively for the Uniform, Cerina and Sobol layouts. In other words, the Sobol layout achieves on average the lowest mean distance between points but at the same time it guarantees the largest minimal separation.

It can be said that the Sobol layout is our response to the Uniform model because it is uniform in its own right. However, when we thought of the solution for the locations getting arbitrarily closed, we imagined that this model would lend itself well to be stacked with different levels that are denser, and this would serve as a model for a monocentric city. By contrast, the Correlated Group Membership model is polycentric, and as such it differs significantly from our own proposal.

Recently, we found out about the diffusion-limited aggregation (DLA) model which is used to model cities [35]. It is different from the Sobol layout (and actually from the Uniform and CGM models) in that the latter aims to place discrete locations because this is what we employ inside the human mobility models. The mental picture to keep in mind is that of disconnected towns surrounding the suburbs and a dense (though still discrete) urban core at the centre; the DLA model aims to capture something that is aggregated and hence contiguous and that we would need to discretise to use in the mobility models.

Distribution of the weights

Having placed the nodes according to our preferred layout, we consider independently all the edges. The probability that node i is connected to node j is set to

$$p_{ij} = \begin{cases} \frac{1}{Z}\lambda_{b_i b_j} f(d_{ij}) & \text{if } i \neq j\,, \\ 0 & \text{otherwise,} \end{cases} \tag{4.16}$$

where $\mathbf{\Lambda} = (\lambda_{rs})$ is chosen to generate the desired connectivity and the normalisation constant Z is set so that $\sum_{i,j} p_{ij} = 1$. We exclude self-edges from being drawn because the human mobility models that our method relies on cannot account for them. In the definition above, we recognise the gravity affinity function, Equation (3.11), with the attractiveness and the propulsiveness being the same for the nodes that belong to the same group.

We generate a multi-graph in which we place a total of $M = \lceil \rho N(N-1) \rceil$ edges (divide this quantity by two for an undirected model) allowing for repeated edges between the same pair of nodes. Once the edges are drawn, we reinterpret the parallel edges in terms of an edge weight. We remark that the weighted density ρ can be larger than one and it should not be confused with the topological density ρ_t which measures the proportion of nodes that share a link out of the total number of possible pairwise combinations (see Section 2.1).

Both Expert and Cerina consider M to an expected value and the number of edges placed is a random variable that fluctuates around it. Since we have normalised the probabilities and they sum to one, we find it easier to draw the edges according to the multinomial distribution with parameters M and $(p_{ij})_{i,j=1,\dots,N}$. In this way the networks that we generate have *exactly* M edges.

The placement of the edges according to p_{ij} creates directed networks, but we can turn this easily into an undirected model by only considering the subset of edges between pairs of nodes that satisfy $i < j$. When the model is directed, in the situation that the mixing matrix $\mathbf{\Lambda} = (\lambda_{rs})$ is symmetric, we expect to see on average the same number of edges between the two groups (no net inter-group flow), though the edges are still drawn independently and the networks sampled will normally have a nonzero net flow due to random fluctuations.

In the model by Expert et al., the deterrence function is $f(d) = 1/d$ and the mixing matrix is

$$\mathbf{\Lambda} = \begin{pmatrix} 1 & \lambda \\ \lambda & 1 \end{pmatrix}, \tag{4.17}$$

with $\lambda \in [0, 1]$. For this range of values, the groups will be connected assortatively if $\lambda < 1$. In our own testing framework, we employ primarily the mixing matrix above but we allow the parameter λ to vary in the interval $[0, 2]$ so as to also generate disassortative block structures for $\lambda > 1$. We leave the testing with more general structures such as core-periphery — which the mixing matrix framework could naturally accommodate — for a future study.

Our own preference for the deterrence function is $f(d) = d^{-\gamma}$, where the parameter γ is our own addition and is helpful to tune the effect of spatial constraints: if $\gamma = 0$, the position of the nodes is irrelevant and the synthetic network is in effect a regular SBM; Expert et al. set $\gamma = 1$, and we are interested to explore the choice $\gamma = 2$ because this value is closer to what we obtain for the empirical fitting of the gravity model applied for example to the network of commuting and the retail network (see Section 3.6).

Cerina's model connects nodes in proportion to $\exp(\beta s_i s_j - d_{ij}/\ell)$ which can be separated

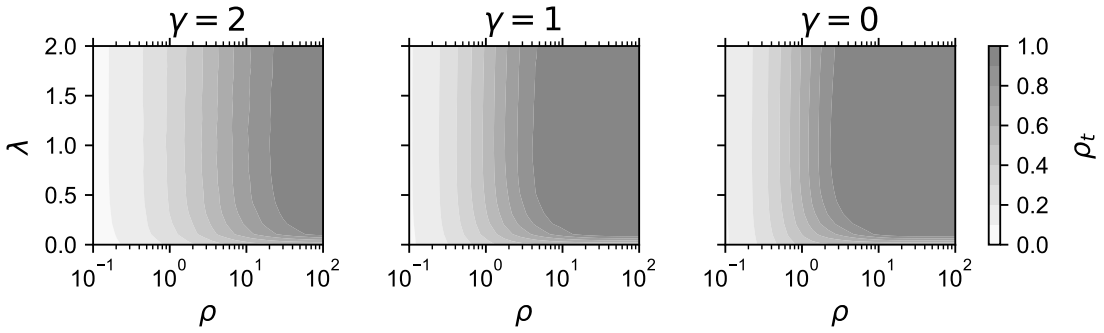


Figure 4.2: Topological density of the (ρ, λ) phase space for different values of γ. The underlying networks were generated with a Sobol layout with parameters $R = 3, p = 3$ resulting in $N = 80$ nodes.

into the deterrence function $f(d) = e^{-d/\ell}$ and a mixing matrix

$$\mathbf{\Lambda}' = \begin{pmatrix} e^{\beta} & e^{-\beta} \\ e^{-\beta} & e^{\beta} \end{pmatrix} . \tag{4.18}$$

Compared to the matrix defined in Equation (4.17), $\mathbf{\Lambda}'$ also has a single parameter, but β controls the inter- and intra community rates at the same time: the planted blocks are undistinguishable when $\beta = 0$ and the limiting $\beta \to \infty$ gives us (after normalisation with the term $1/Z$) the perfectly assortative block structure that we can also produce with $\lambda = 0$ inside $\mathbf{\Lambda}$.

We have discussed the different choices for the layout of the nodes which we will explore in turn. We now describe the experiment that will let us compare the layouts as well as the different choices that we can make inside the space-corrected SBM.

In summary, once we have chosen the layout, there are three parameters that are left to control the sampling of the weights: $\rho \geq 0$ tunes the magnitude of the weights generated and indirectly the density of the network; $\lambda \in [0, 2]$ controls whether the blocks are connected more assortatively ($\lambda < 1$) or disassortatively ($\lambda > 1$); and $\gamma \geq 0$ calibrates the strength of spatial constraints.

We look at the (ρ, λ) phase space for fixed values of γ. We vary ρ on a logarithmic scale spanning $[0.01, 100]$ and we linearly sample $\lambda \in [0, 2]$. As mentioned before, we select $\gamma \in \{0, 1, 2\}$ as a representative sample of the varying extent of spatial constraints.

We manage all of our experiments with the excellent Python library Signac [2, 96], a data management framework that lets us specify an abstract parameter space and run repeated recovery experiments at each point in this space. Between the different combinations of model parameters and choices for the SBM algorithm, we sampled over one hundred thousand points of the parameter space, and this task would have been much harder without Signac.

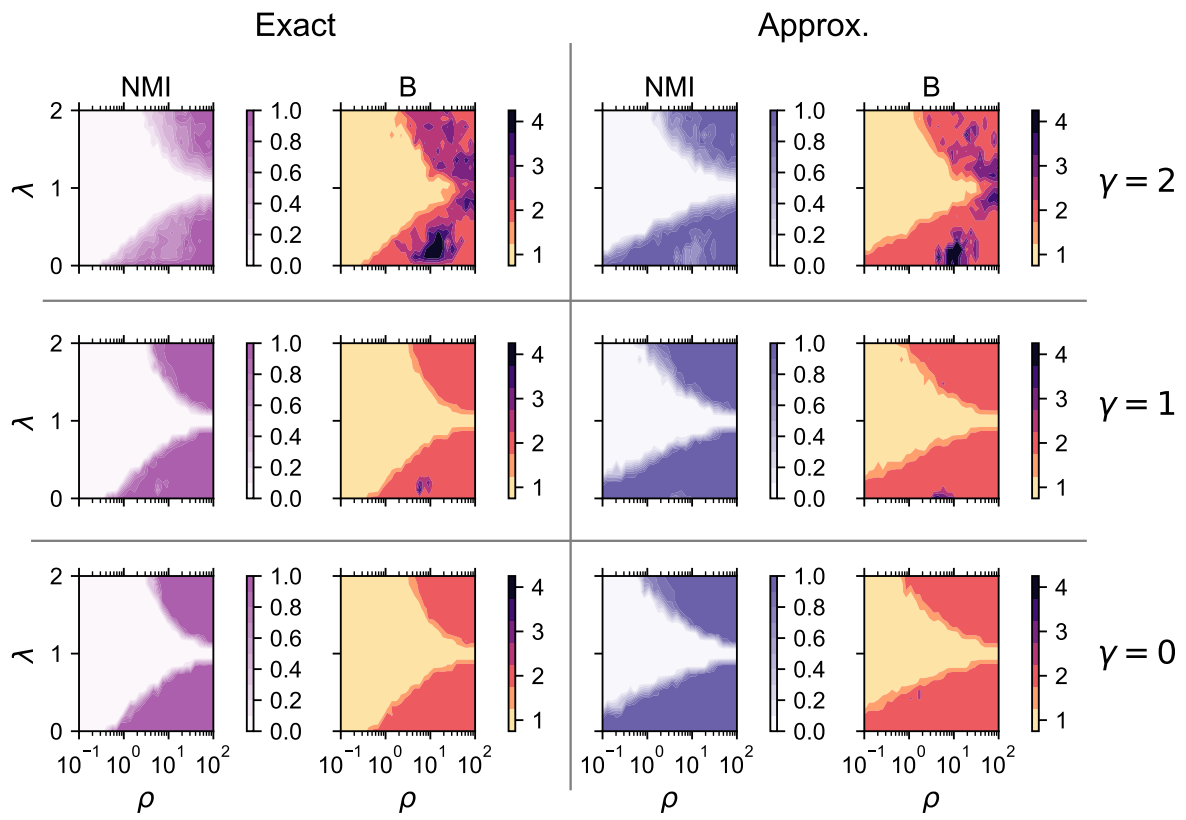


Figure 4.3: Phase space of the Sobol layout for the doubly-constrained gravity model and a layered SBM, as measured with the average NMI of the recovered partition and the average number of blocks. The p-values are calculated using either the exact binomial test (left) or the approximate Z-test (right). For each (ρ, λ) pair, we generated 10 networks and we repeated the SBM fit 5 times, for a total of 50 repetitions.

4.4.2 Recovery results for the Sobol layout

On to the experiments. We start with the Sobol layout with $R = 3$ levels and $p = 3$, resulting in $N = 80$ nodes, the same parameters that we employ to generate Figure 4.1B. In Figure 4.2, we plot how the topological density ρ_t of the synthetic networks varies across the (ρ, λ) phase space for different values of γ. If ρ is high enough, we find that the topological density tends to 1 (as long as $\lambda > 0$, otherwise nodes in different blocks cannot be connected). However, the value of ρ at which the networks display full density increases significantly with γ; we see this in terms of the shrunk grey region when $\gamma = 2$ and a less dramatically when $\gamma = 1$. This makes sense because a larger γ translates into stronger spatial constraints that prevent nodes that are that are too far apart from being linked due to a vanishing small probability $p_{ij} \propto d_{ij}^{-\gamma}$.

Effect of the spatial constraints and the human mobility model

In Figure 4.3, we apply the layered SBM to the signed backbones derived from the doubly-constrained gravity model. We employ either the binomial or Z-test to establish the significance of edges and construct the spatial backbone. The average normalised mutual information (NMI) is calculated between the planted blocks and the partitions recovered using our algorithm across

multiple runs; the darker the colour is, the higher the NMI and the closer that our method gets to recovering the ground truth. We also find it useful to keep track of the average number of blocks B in the recovered partitions.

In each panel in Figure 4.3, that is for different values of γ and for the two types of significance test, we observe essentially the same behaviour: there are two regions on either side of $\lambda = 1$ where recovery of the planted groups is possible (for brevity we refer to these as the feasible regions). Setting $\lambda = 1$ makes the two groups statistically indistinguishable of each other and is therefore a natural detectability barrier that will affect any recovery method. However, we are happy to report that the space-corrected SBM can recover (to a varying degree depending on the experiment) the ground truth both for assortative and disassortative synthetic networks. Furthermore, if we focus on an individual panel and we take vertical cross-sections, we also find that the region where recovery is feasible expands monotonically with respect to the edge density ρ. Outside of the feasible recovery regions, where the statistical evidence is lacking, the number of blocks is equal to 1.

In addition to these general trends, we can compare the panels in Figure 4.3 from top to bottom and conclude that the recovery problem becomes easier when the spatial effects are made weaker with smaller values of γ. The limiting case $\gamma = 0$ gives us a synthetic model where the generated weights only depend on the block membership of the nodes — in other words, our model is equivalent to a simple SBM — and it is clear that our recovery algorithm which itself employs an SBM will work well in this case. That being said, it is still interesting to find that the density ρ plays the role of expanding the feasible region. We remember that, by construction, the signed backbone recovers the signals of the positive and negative edges which can be quite sparse, therefore the recovery is affected by densities that are too low to begin with.

If we now compare the left and right panels, we also find that the Z-test approximation leads to an improvement seen in the fact that the feasible region expands to the left side of the plots corresponding to lower densities. This is due to the fact that the Z-test approximation is less stringent on the edges that make it to the signed graph, particularly for positive edges, and the result is that the recovery is slightly less impacted by the lower densities.

Since the sampling of the weights in our synthetic network models is done using a power-law deterrence function, it seems reasonable to expect that the gravity model will lead to better recovery results than the radiation model. Indeed, we remember that the assumption underlying the construction of the spatial backbone is that the selected human mobility model is a modelling how the network came to be. By nature of the sampling of the weights, we have that the gravity law is naturally closer to the real mechanism that created the synthetic networks than the radiation model.

Let us now examine the recovery based on the radiation model, as seen in Figure 4.4. Since the production-constrained model is almost as good as the doubly-constrained one but is also much cheaper to construct, we use it in our experiments.

The phase space looks nothing like it did before for the gravity law. The gap between the feasible regions on either side of $\lambda = 1$ has grown wider and the recovery is very poor for disassortative block structures (at least in the range of our experiments with $\lambda \leq 2$). In

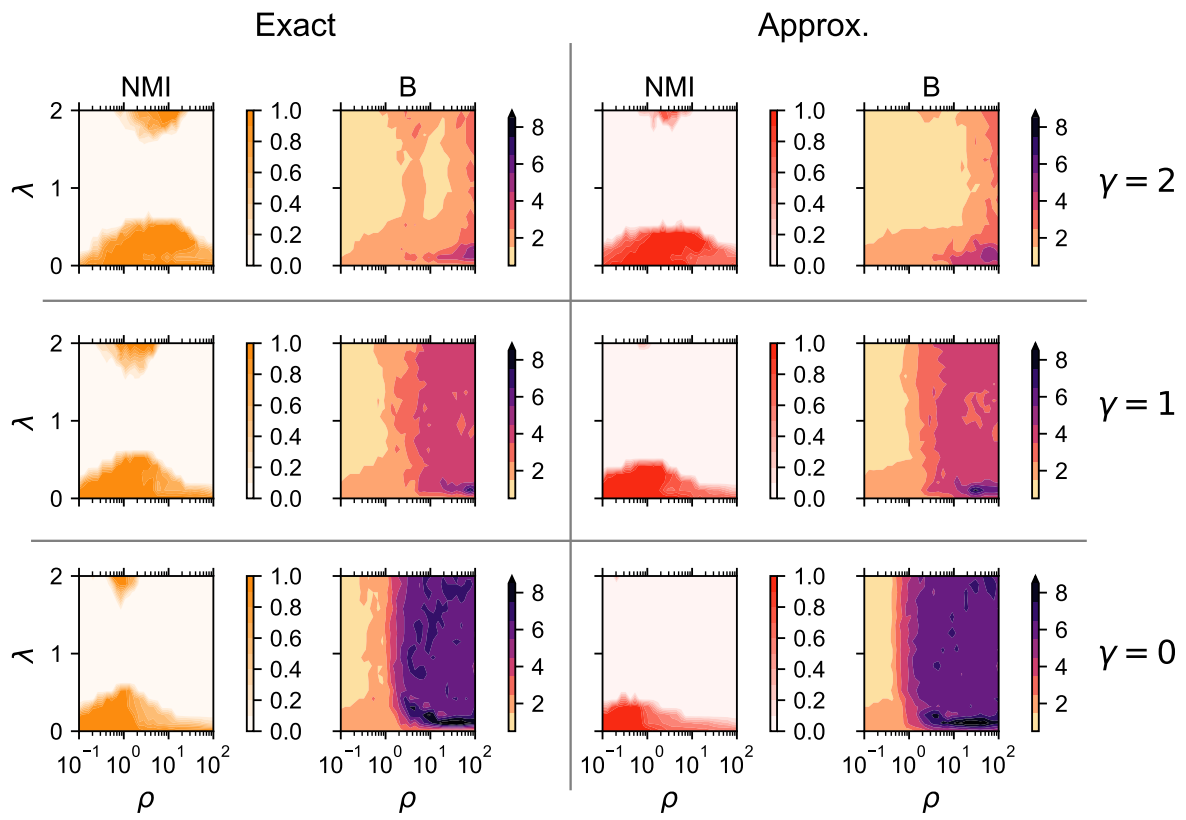


Figure 4.4: Phase space of the Sobol layout for the production-constrained radiation model and a layered SBM. The p-values are calculated using either the exact binomial test (left) or the approximate Z-test (right). For each (ρ, λ) pair, we generated 10 networks and we repeated the SBM fit 5 times, for a total of 50 repetitions.

the assortative region $\lambda < 1$ the portion of the phase space with high NMI no longer grows monotonically with respect to ρ; in fact, a high density degrades the recovery region with respect to λ. This is explained by the fact that number of blocks increases very significantly with ρ, but we will have to wait until the next paragraph to find the reason for this increase.

Comparing the panels from top to bottom, we obtain an increasingly worst recovery if γ decreases, exactly the opposite behaviour to that of the gravity law. The radiation model is based on the matrix of intervening opportunities. If $\gamma = 0$ and the weights generated do not depend on the distances, any node can be connected to any other node and there is no notion of intervening opportunities; if, on the other hand, $\gamma = 2$, the rapid decay of the term ${d_{ij}}^{-2}$ more closely resembles the situation being modelled with with intervening opportunities.

For comparison, the same set of recovery experiments when we use the Spatial modularity of Expert et al. [27] can be found in Appendix A, Figure A.1. The method obviously fails to recover anything if $\lambda > 1$ but it also fails to recover a large portion of the assortative region unless $\gamma = 0$. When the recovery fails, the modularity-based method also overfits the synthetic network finding as many as 8 different groups on average.

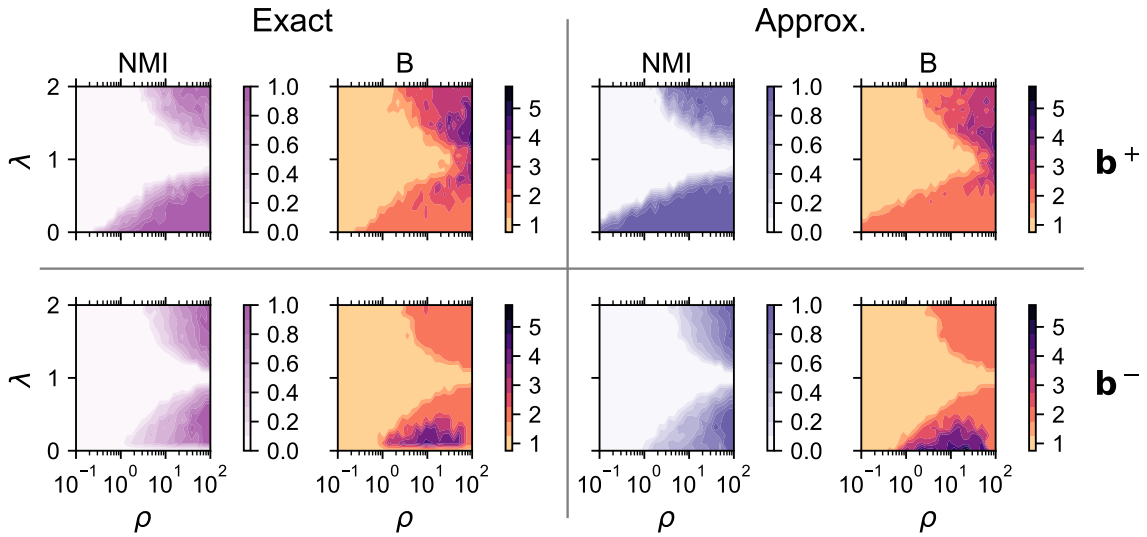


Figure 4.5: Phase space of the Sobol layout for the doubly-constrained gravity model applied to the positive and negative layers independently. The power-law parameter is set to $\gamma = 2$. For each (ρ, λ) pair, we generated 10 networks and we repeated the SBM fit 5 times, for a total of 50 repetitions.

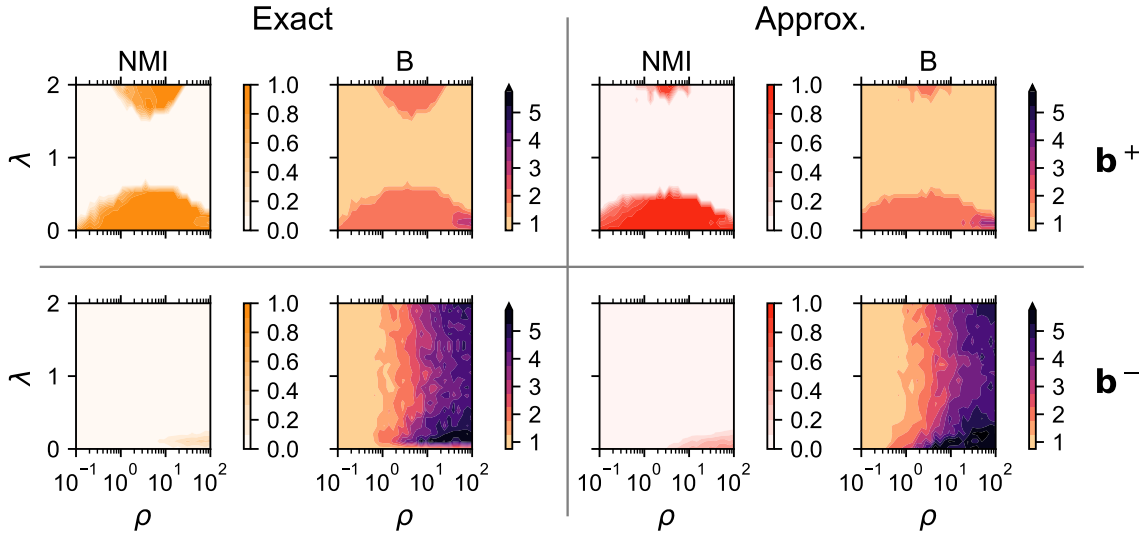


Figure 4.6: Phase space of the Sobol layout for the production-constrained radiation model applied to the positive and negative layers independently ($\gamma = 2$; 50 repetitions for each point).

Separate signals in the signed networks

In Figure 4.5, we investigate the recovery based on the DC gravity model and a monolayer SBM fit to the positive and negative networks separately. In the interest of space, we only examine the results for fixed $\gamma = 2$ which, as we saw previously, is the value leading to the hardest recovery

problem.

The recovery using the positive partition $\mathbf{b}^+$ (top panels) is overall very similar to the layered SBM results and therefore quite satisfactory. We observe, however, two areas where the two sets of results differ: when the exact binomial test is employed, the positive partition $\mathbf{b}^+$ finds $B = 2$ for the assortative networks more often than the layered partition; when using the approximate Z-test, the layered partition finds $B = 2$ more often that its signed counterpart.

On the other hand, the results obtained for the negative partition $\mathbf{b}^-$ (bottom panels) are not as satisfactory because the feasible regions are clearly shrunk compared to the layered partition (and the positive partition). As expected, the negative backbone cannot deal with the perfectly assortative networks as attested by the light narrow band at the bottom of the exact NMI figure that was obtained for the value $\lambda = 0$. However, a positive finding is that $\mathbf{b}^-$ very clearly finds that the disassortative region is best partitioned into $B = 2$ blocks (even if a sizeable portion of the nodes is actually mislabelled because the NMI is not equal to 1). The layered SBM which combines information from both the positive and negative signals at once is therefore doing a better job than either one of the two signals handled separately.

Turning our attention to the radiation model in Figure 4.6, we now find that the results obtained with the positive partition $\mathbf{b}^+$ do not look like the layered partition. The feasible regions are still shrunk compared to the gravity law, but inside them we get a perfect recovery with an NMI of 1 and $B = 2$ blocks. On the other hand, we cannot recover anything of value with the negative partition and the number of blocks increases from a single block if the density is lower than $\rho = 1$ to as many as 5 or 6 blocks on average if ρ is increased further. This behaviour matches exactly what we obtained with the layered SBM and thus we conclude that the radiation-based layered SBM is impacted by the signal due to the negative edges in $\mathbf{A}^-$.

4.4.3 Uniform and Correlated Group Membership models

In Figure 4.7, we present the recovery results for the Uniform and Correlated Group Membership models. We first analyse the method based on the layered SBM and the doubly-constrained gravity law, the combination that led to the best results for the Sobol layout. The densities and the results obtained using the radiation model can be found in Appendix B. We employ the power law deterrence function with fixed $\gamma = 1$.

The recovery results for the Uniform model are excellent (top rows in Fig. 4.7), though this is true for easier parameter of $\gamma = 1$ and the results are degraded for the harder $\gamma = 2$ (we do not include these results for reasons of space). The recovery of the planted blocks for uniformly spaced nodes is sensitive to points that lie too close to each other. Indeed, in our experiments we found that it is not uncommon to generate a network in which a single pair of points lie so close that p_{ij} and p_{ji} account for more than half of the total of the summed probabilities (a problem that is exacerbated when we square the distances). This in turn leads to two edges that are important outliers in terms of their associated weight — remember that we are interpreting parallel edges as weights — and therefore a very sparse positive backbone. This situation is what the Sobol layout was designed to avoid.

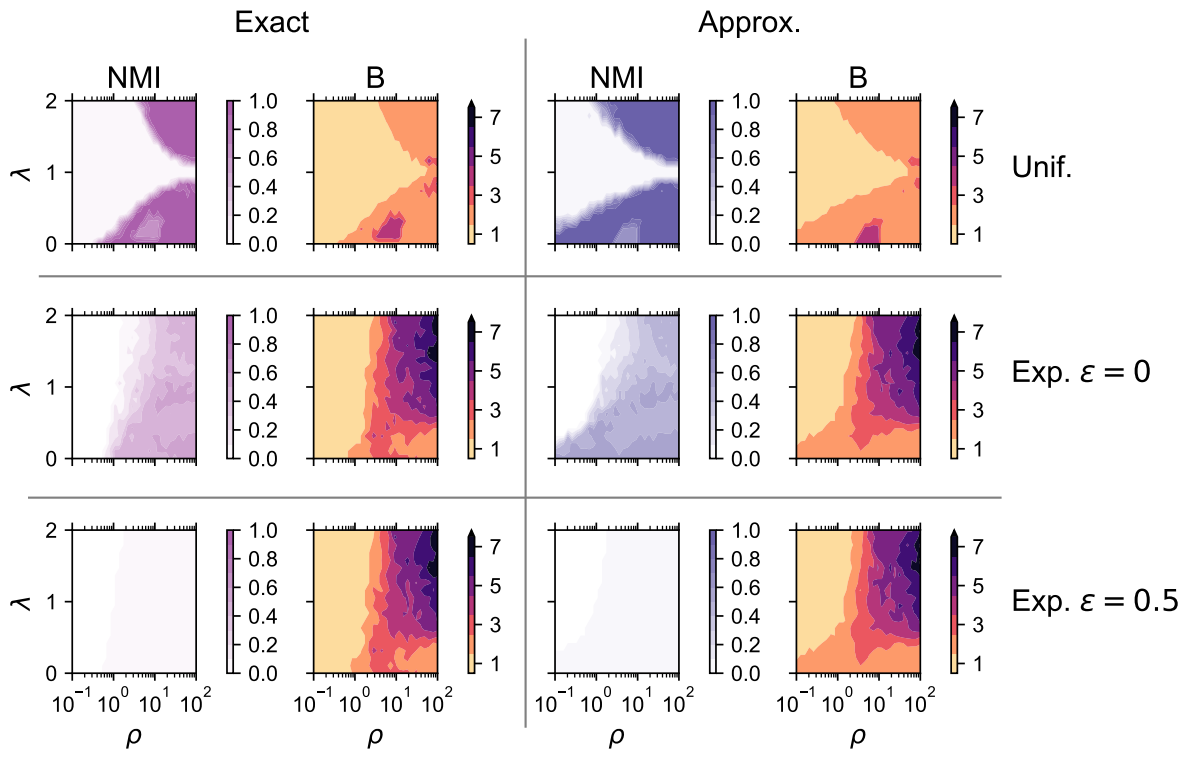


Figure 4.7: Phase space of the Uniform (top) and the Correlated Group Membership layouts, with perfect spatial correlation (middle) or no correlation (bottom). The weights are generated with the power law distribution with coefficient $\gamma = 1$ and the recovery employs the doubly constrained gravity model.

Perhaps more interesting are the very poor results that we obtain when the nodes are placed according to an exponential distribution around two centres, the layout of Cerina et al. If the group membership is perfectly correlated with space via the parameter $\epsilon = 0$ (that is, all of the nodes on either side of $x = 0$ belong to the same group) we get an NMI of around 0.5 only if the density is high enough. However, for the uncorrelated situation with $\epsilon = 0.5$ the NMI is only as high as 0.03. In both cases the average number of blocks can be as high as 7 if the density is above $\rho = 10$.

The Correlated Group Membership model is also affected by nodes that lie too close to each other, though the main culprit of the terrible recovery is a different mechanism. Cerina et al. proposed a synthetic model that has an exponential deterrence function, and this is well suited to the exponentially distributed nodes. In our experiments with the power law deterrence function, the term $1/d$ (or worse $1/d^2$) decays too quickly given the scale of the system and essentially disconnects the network. We refer the reader back to Figure 4.1 where we compare the layouts: even though the three panels were generated with the scale parameter $\ell = 1$, the exponential layout specified by the CGM model covers an area that is roughly 50 times larger and the histogram of pairwise distances has a much longer tail.

Instead, we change the sampling of the weights and select the exponential deterrence function $f(d) = e^{-d/\ell}$ like Cerina et al. originally do. We also employ the mixing matrix $\mathbf{\Lambda}$ parametrised by β, which we vary over the interval $[0.1, 10]$. We see the new recovery results in Figure 4.8,

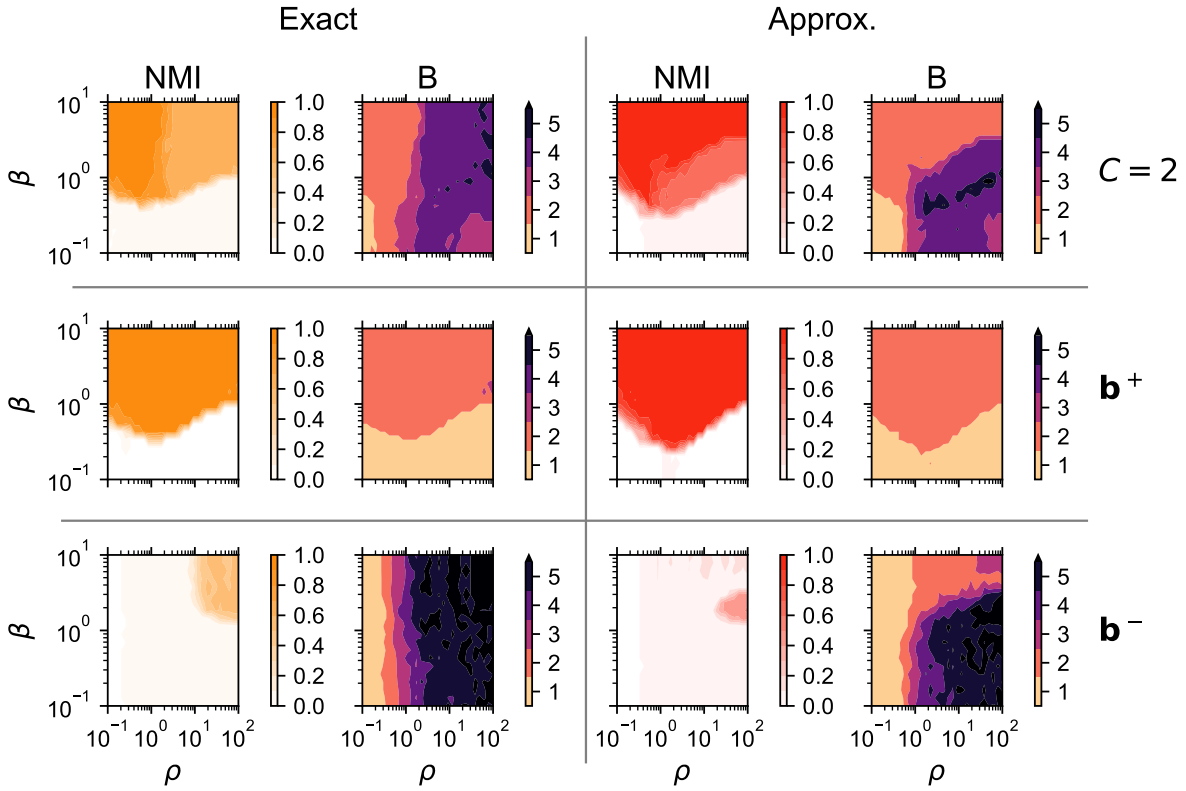


Figure 4.8: Correlated Group Membership layout with exponential weights. The block assignment is not correlated with with space $\epsilon = 0.5$. We obtain the best results for the radiation model (pictured) and in this case the positive backbone.

which was obtained for the uncorrelated case $\epsilon = 0.5$ with the layered SBM (top panels) or the signed networks separately (middle and bottom panels). This example was obtained with the radiation model that leads to best results, and the gravity model (and also the correlated case, $\epsilon = 0$) can be seen in Appendix B.

The recovery based on the positive partition $\mathbf{b}^+$ is very good, with perfect recovery being possible for all the densities explored if $\beta > 1$, even $\beta > 0.5$ if $\rho \approx 1$. The number of blocks found outside of this feasible region is exactly $B = 1$ indicating that there is no statistical evidence to distinguish the connection rates of the nodes. The negative links however are not detecting a signal that is strong enough to do any sort of recovery, and the result is that the layered SBM recovery is affected by the negative edges and therefore performs worse than the positive backbone on its own.

The gravity model (Figure B.5 in Appendix B), on the other hand, does not show the same type of sharp transition because the NMI is never very low nor is it ever much higher than 0.6. The number of blocks is relatively high indicating that the SBM is overfitting the data.

We said earlier that the point of the model by Cerina et al. is to test different regimes of spatial correlation of communities. The authors classify the different regimes and they postulate which method should and should not work: if the links are mostly between neighbouring nodes ($\beta \ll 1$), the case correlated case $\epsilon = 0$ is easy because the attribute communities also correspond

to the spatial communities and any community detection method should work; the uncorrelated case $\epsilon = 0.5$ is hard because the few links between nodes that belong to the same community can be anywhere in space and the method needs to remove the spatial effects in order to find them. If the spatial component of the links is irrelevant ($\beta \gg 1$) any community detection method should work.

This is not the behaviour that we find our method to have. The gravity model has an erratic behaviour and generally overfits the data. This is due to the power law (the one used inside the backbone construction) being a bad fit to the exponentially distributed weights. The radiation model, on the other hand, is well suited to describe these synthetic networks but our method does not distinguish particularly between the correlated and uncorrelated regimes and it discriminate instead between $\beta \ll 1$ and $\beta \gg 1$. In the former case, our method finds just one community. This means that the spatial backbone correctly removes the effect of space, and according to our method there is no other signal left in the data. Thus, our method puts all the nodes in the same group. This behaviour is not entirely undesirable if the attributes are correlated with space, but it is not what we would like in the uncorrelated case. It means that for $\epsilon = 0.5$, the signal left due to the nodes with the same attributes is not strong enough after removing the effect of space (which makes sense because space was the dominant factor).

4.4.4 Analysis of the validation

We have investigated the behaviour of the space-corrected SBM in controlled experiments and found that our method works in a variety of situations but it also fails in different settings.

There is a natural detectability threshold when the combination of parameters make the planted groups statistically indistinguishable.[4] In these situations, the recovery based on the gravity model seems to find $B = 1$ blocks most of the time so we can consider this a success. This behaviour is made possible because we base the detection on a statistical model and we believe that this is certainly one of the most positive aspects of making use of SBMs for community detection.

Between the layered SBM and the positive backbone — and selecting either the gravity or the radiation model — we seem to find feasible regions for the three different layouts. We ask ourselves if either one of them bears some resemblance to the local authorities or the stores that we studied previously and therefore whether we should prefer a given mobility model or the layered SBM, etc.

Generally speaking, there is no simple answer to this question because the synthetic models are just toy models and they each fail to capture different aspects. Considering only the location of the nodes, the Uniform and Sobol locations are probably not a good model if the scale of the system being modelled covers the whole of the UK, but the latter might still be a somewhat realistic toy example of large metropolitan areas in general and London in particular. On the other hand, the centres and the exponential distribution could resemble a larger scale with different metropolitan areas, so we can imagine that a real dataset would be somewhat of an

[4] That being said, we are lacking theoretical guarantees about where the threshold lies and we cannot know for sure if there is still room to expand the regions where recovery is possible.

interpolation of these models.

The most likely scenario is that the gravity and the radiation models will highlight different aspects of a real-world spatial network and in the following sections we will study them both. The validation has shown that we can be confident that the space-corrected SBM will recover the signals that are strong enough.

Furthermore, in Chapter 3, we mentioned that for the gravity law, the power-law deterrence function is a logarithmic re-scaling of the distance compared to the exponential function. The drawing of the weights in the validation we just did showed precisely that the two choices have different purposes depending on the size of the system. However, the recovery that we use is only based on the power-law function and it will be important in future work to include an exponential deterrence function.

The last point we make here is that we did not explore the effect of the significance level α which controls the density of the spatial backbone. Given that there are already a great deal of choices to be made we found it easier to leave this parameter fixed for the time being, but it would certainly be of great interest to do this exploration in the future.

4.5 Space-corrected analysis of the commuting network

We return to the spatial network defined by commuting in the UK as inferred from the 2011 UK census. The networks is composed of 404 nodes, one for each local authority. The weighted and topological densities of this network are respectively $\rho = 69.1$ and $\rho_t = 0.664$ so, according to the validation done previously, we should be in a regime that is dense enough to detect meaningful groups of nodes.

The scatter plots in Figure 3.5 give us a rough idea of the fit obtained with the doubly-constrained gravity model and the doubly-constrained radiation model, and the resulting edges (in colour) classified as positive or negative in the signed backbone. For the gravity law, the negative edges that are considered significantly low according to an exact test outnumber the positive ones roughly three to one as we can read off of Table C.1 in Appendix C; for the radiation model on the other hand, the situation is reversed and the positive edges outnumber the negative ones roughly twenty to one.

This is as good a place as any to remind the reader that positive edges are indicative of observed commuting flows that are larger than expected (while negative edges signal flows smaller than expected). In the context of stochastic block models, we will be looking for groups of nodes with many positive (negative) edges and this will then tell us about regions that are interacting more strongly (weakly) than expected.

4.5.1 Gravity partitions

We start by comparing the independent positive and negative partitions that were based on the DC gravity model in Figure 4.9 and we will then proceed to the results for the layered SBM.

Each partition is the overall best with minimal entropy out a hundred different runs of the DC-SBM. The two point estimates found partition the positive or negative backbones (and thus

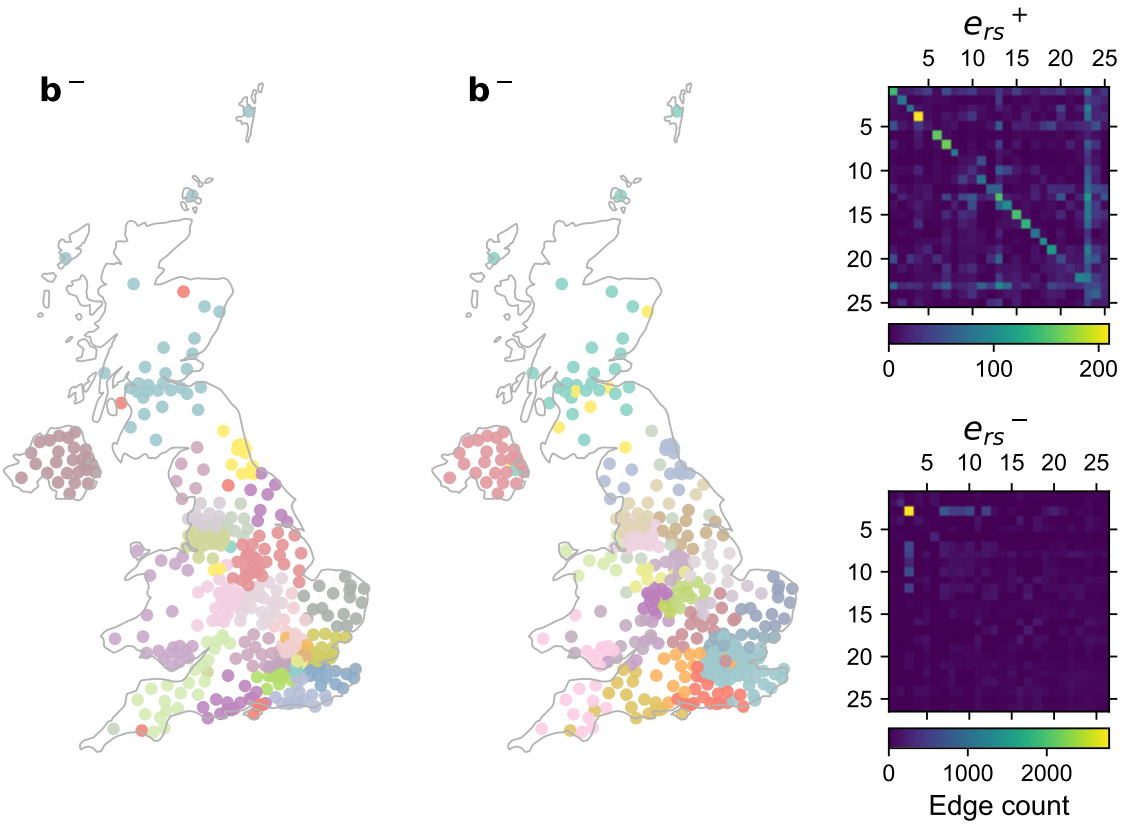


Figure 4.9: Minimum description length partitions and edge count matrices for the positive and negative backbones, as based on the doubly-constrained gravity model.

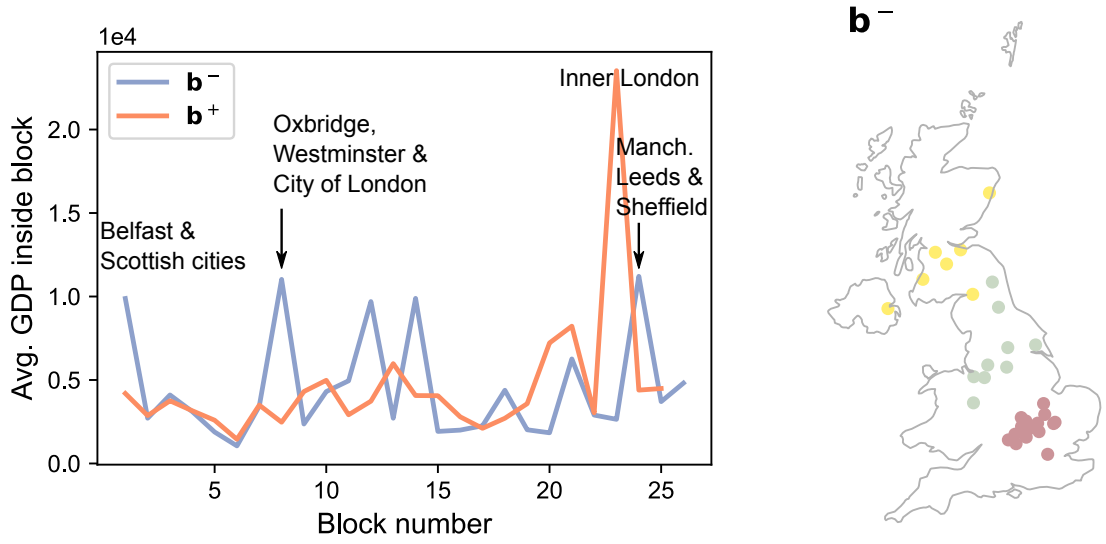


Figure 4.10: (Left) Average per capita GDP computed for the nodes assigned to a given block. We have annotated some interesting groups whose composition we can interpret. (Right) Highlighted blocks in the negative backbone.

different networks) respectively into 25 or 26 blocks. The agreement between the two partitions is $\mathrm{NMI}(\mathbf{b}^+, \mathbf{b}^-) = 0.665$ which is relatively high, reflecting that there are groups of nodes that appear in both partitions like Northern Ireland, Scotland, the region around Liverpool and Manchester , and to some extent Lancashire (located just above Manchester) and Norfolk. That being said, there are important dissimilarities too, like the negative partition breaking up Wales into three separate groups and the positive one identifying it as a whole group (this in turn is an interesting finding because our method is picking that people commute more than expected within the country, a signal that is definitely cultural). The region around South-East London is also grouped together in the negative backbone and broken up in the positive one.

Inside the matrix of edge counts corresponding to the positive backbone, $\mathbf{e}^+$, the first thing we notice is that the matrix does not entirely correspond to an assortative pattern, though there are large entries along the diagonal. Some of the largest entries along the diagonal correspond to Scotland, Northern Ireland, Liverpool and Manchester and Kent. Out of these, position 13 in the matrix which corresponds to Liverpool and Manchester shows a large influx from the communities 14 and 20, the first one roughly corresponding Hertford and Bedford and the latter roughly corresponding to Berkshire and West London. That is, we have an unexpectedly large number of commuters travelling from two areas surrounding London and coming to work in the Liverpool and Manchester area (and looking at individual pairs of nodes, the majority of the incoming flows are directed to Manchester in particular).

The previous result might seem odd and one might see a contradiction because we could reasonably expect the opposite conclusion, that a large number of residents of Manchester commute into London. The apparent contradiction is resolved because we are confusing two different things: the magnitude of the flows and the number of edges that are deemed statistically significant. Indeed, there is a much larger number of commuters from the areas around Liverpool and Manchester that work in London that the other way around. However, a good portion of these flows are being accounted inside the gravity model and they are not considered surprising. On the other hand, a large number of flows from London and into Manchester are statistically significant even if overall they are smaller in magnitude.

It is by design that the spatial backbone places the same importance on every edge that is significant (irrespective of magnitude) because this lets us employ the unweighted stochastic block model. Nevertheless, it is true that as a result we might obtain a noisier signal — or at least a signal that is harder to interpret, like we had above — and in the future we would like to study the extent to which this can be remedied if we vary the significance threshold α (currently fixed and equal to 0.01).

Going back to the analysis of the $\mathbf{e}^+$ matrix, we now focus on the column corresponding to community number 23. This community is very well connected to other parts of the network, specially in terms of incoming fluxes (note the bright row and specially the bright column number 23). Community 23 is composed of only nine nodes, all of them in central London: Camden, Hackney, Hammersmith and Fulham, Islington, Kensington and Chelsea, Lambeth, Southwark, Tower Hamlets and Westminster-City of London. The fact that the column is brighter than the row number means that these nine nodes are net receivers of commuting flows. Few countries

are as centralised as the United Kingdom and our method easily picks up this signal.

When we look at the matrix of edge counts for the negative backbone, the bottom matrix in Figure 4.9, the results are very skewed by a single community that has a very high number of internal edges — this is also the largest community with 70 nodes; the second largest community has 28 nodes in it. This is the community in light-blue which surrounds London and extends to the South-East. Some of the nodes in this community are located within the Greater London area (Greenwich, Lewisham, Hackney, Harrow, Hillingdon, Ealing, Croydon) as well as in Kent (Maidstone, Ashford, Tunbridge Wells, Swale, Dover, Gravesham, Sevenoaks, Thanet, Medway) and Sussex (Crawley, Hastings, Lewes, Adur, Rother). Since we are working with the negative backbone, edges in this network are related to observed flows that are significantly lower than expected. The fact that this large entry lies on the diagonal tells us that the group of nodes interact less than expected, even if they are located close to each other. Such a structure around London makes sense because many of the commuting trips among these nodes will be directed to the innermost areas of London. In other words, we have found a group of nodes that has the function of the periphery in a core-periphery mesoscale structure.

We computed the average per capita GDP of the local authorities contained in each block at the time of the census in 2011,[5] and the results are in Figure 4.10. There are no particular outliers for the positive backbone except for community number 23 made of the nodes in central London, but the negative backbone is very interesting. To the right, we have highlighted three groups of nodes in the negative partition with a particularly high average GDP: block number 1 includes Aberdeen, Edinburgh, Glasgow and Belfast (Belfast being the only node outside of Scotland); block number 8 includes Oxford, Cambridge, Milton Keynes and Westminster-City of London; in block number 24 we can find Sheffield, Leeds and Manchester. In the three groups, we recognise some of the most important cities in the different regions of the country. We find it particularly striking that Belfast is grouped with the large Scottish cities revealing a similarity that overcomes the physical barrier of the Irish channel. It is also very interesting that Oxford and Cambridge — university towns but also innovation hubs and political powerhouses — are grouped together with Westminster and the City of London.

It is important to point out that we employed the GDP to find interesting blocks to focus on, but the organisation of the network that led to the partition $\mathbf{b}^-$ in the first place does not depend on the GDP directly, it depends solely on whether the commuting flows were larger or smaller than expected given the location of the nodes. From these examples, we can positively conclude that we have been successful in removing at least some of the spatial effects in order to reveal other aspects, economical or cultural, that that are taking place inside our network.

The results obtained using the layered SBM can be found in Figure 4.11. This time there is a single partition into $B = 27$ groups. This grouping of nodes, is similar with both the $\mathbf{b}^+$ and $\mathbf{b}^-$ vectors that we just described, with the similarity being $\mathrm{NMI}(\mathbf{b}, \mathbf{b}^+) = 0.808$ and $\mathrm{NMI}(\mathbf{b}, \mathbf{b}^=) = 0.795$.

This time, the two matrices in Figure 4.11 are directly comparable because they were obtained

[5] Source: https://www.ons.gov.uk/economy/grossdomesticproductgdp/datasets/regionalgrossdomesticproductlocalauthorities

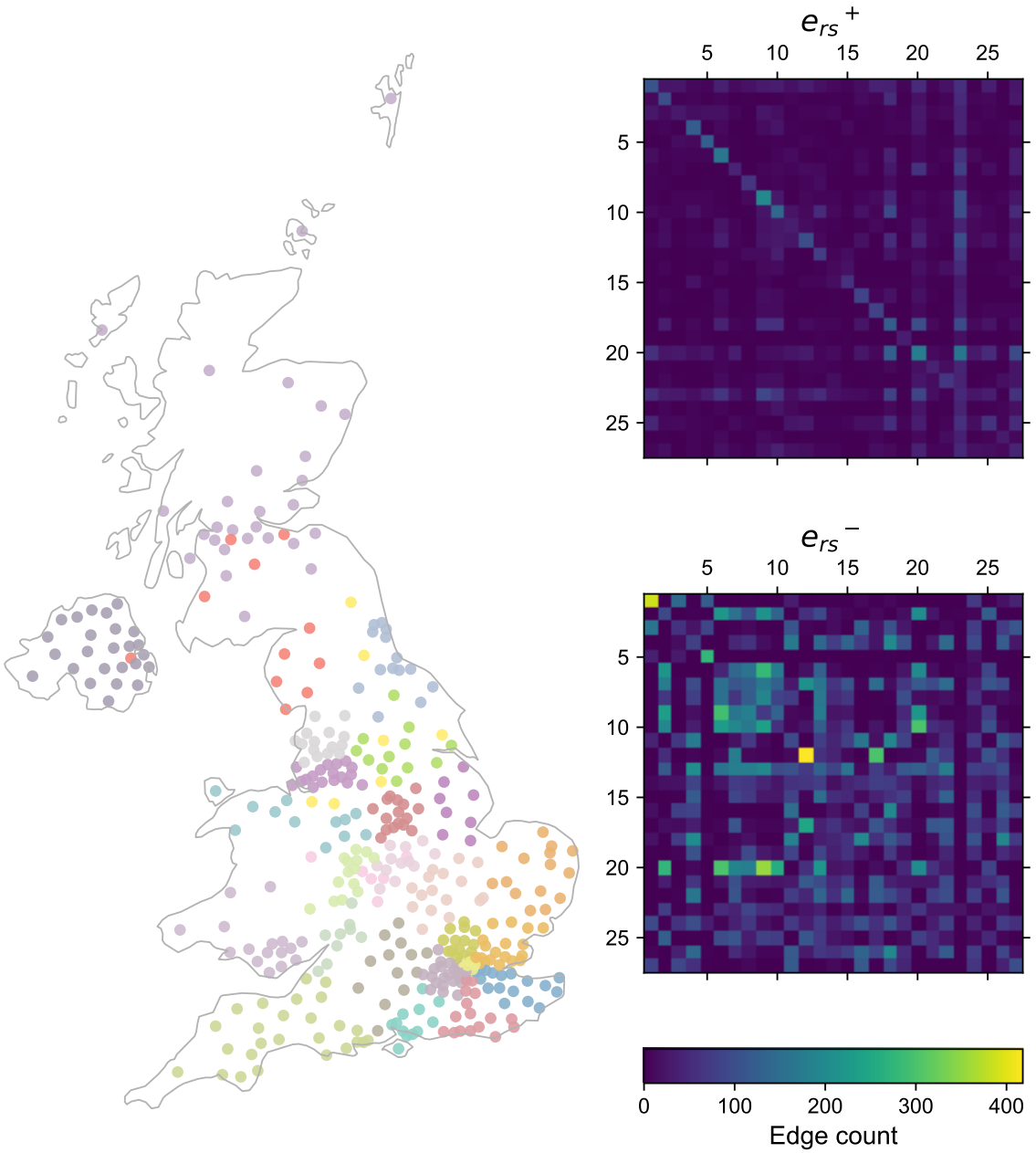


Figure 4.11: Minimum description length partitions and edge count matrices for the layered SBM, as based on the doubly-constrained gravity model.

for the same groups of nodes (and we used the same colour scale to reflect this). We recognise a familiar pattern in the brighter rows and columns inside $\mathbf{e}^+$: blocks 18, 21, and 23 which correspond respectively to Manchester and Liverpool, the are around Bristol and the innermost London (the exact same nine nodes). These metropolitan areas are net receivers of commuting flows.

Turning our attention to the negative edges, the matrix $\mathbf{e}^-$ is perhaps harder to read because it is denser but we highlight some large entries in the diagonal: the four largest entries regions which are significantly less connected than we could expect are Northern Ireland (excluding Belfast; this is block number 5 and we note as well that this block is almost perfectly assortative), the Northern counties in Scotland, Cornwall and the region in South-East London roughly overlapping with the Berkshire group that we discussed before.

Outside of the diagonal elements, blocks 2, 6, 7, 8, 9, 10 , 13, and 20 form a dense group of tightly linked blocks indicating that they are under-connected given their proximity (we highlight them inside the leftmost panel in Figure 4.12). All of these blocks are in South-East England with the exception of the light pink group corresponding to Birmingham and the Midlands. Clearly this nodes are poorly connected amongst themselves because they surround London.

Also in Figure 4.12, we investigate the average GDP of the blocks. We find the usual regions: the inner core of London Boroughs, the Scottish cities with Belfast, the metropolitan area around Manchester and Liverpool, the Midlands and Birmingham. This time Oxford is not grouped with Cambridge and Westminster-City of London, in fact Cambridge is placed in the same block as Norfolk and therefore does not rank very high according to average GDP. In this separation, the layered partition is more similar to $\mathbf{b}^+$ but in the block that groups Belfast together with Edinburgh and Glasgow the layered partitions borrows more from $\mathbf{b}^-$. All in all, the three partitions are interesting when viewed together.

4.5.2 Radiation partitions

The partitions we obtain using the doubly-constrained radiation model can be found in Figure 4.13. In the validation with synthetic networks, it made sense to employ to the less precise but computationally cheaper singly-constrained model because of the large parameter space we needed to explore but here we can construct the spatial backbone using the doubly-constrained model that provides a better fit.

As before, we repeated the SBM fit a hundred times and kept the lowest entropy estimates. The positive partition has 12 blocks and the negative one has 18. Their agreement as measured using the NMI is 0.492 and, visually, we find that the two partitions start to disagree more than they agree. Similar to what happened with the gravity model, we find that the negative backbone has mainly an assortative pattern as indicated by the bright diagonal in the matrix of edge counts $\mathbf{e}^-$, while there is richer organization according to the positive matrix $\mathbf{e}^+$.

The structure of the positive backbone is further investigated in Figure 4.14. In this schematic division we clearly distinguish several cores: Central London, Manchester and the Midlands, and the stretch of Scotland linking Glasgow to Edinburgh. Surrounding these cores, we find peripheral blocks that are more or less concentric, and then we find the three most peripheral

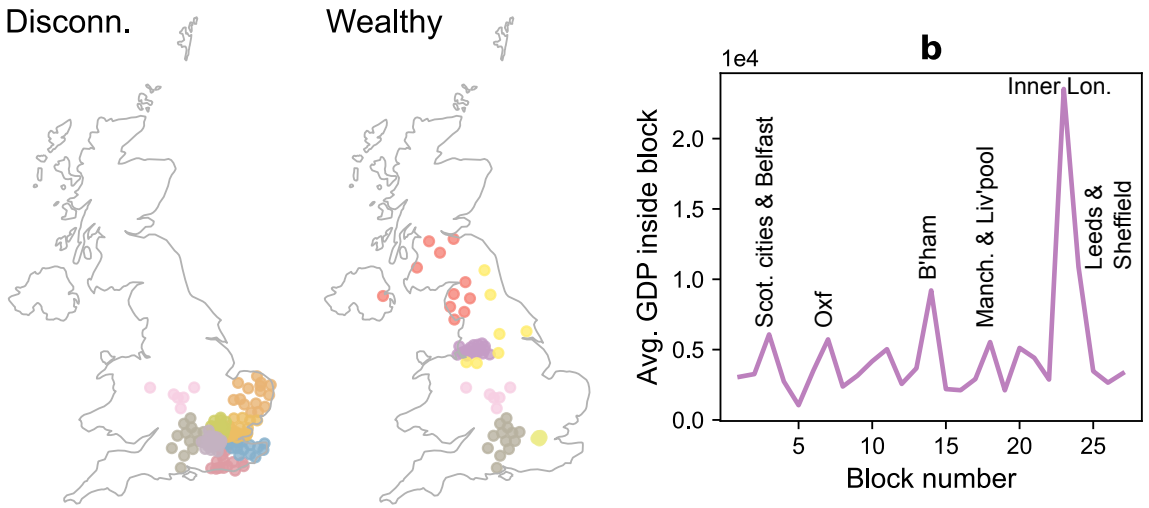


Figure 4.12: (Left) Group of eight blocks that forms a dense cluster in the negative backbone and thus indicates weakly linked nodes. (Centre) selection of the top six wealthiest regions as measured by the average per capita GDP (right). The underlying partition was obtained with the layered SBM and the DC-gravity model.

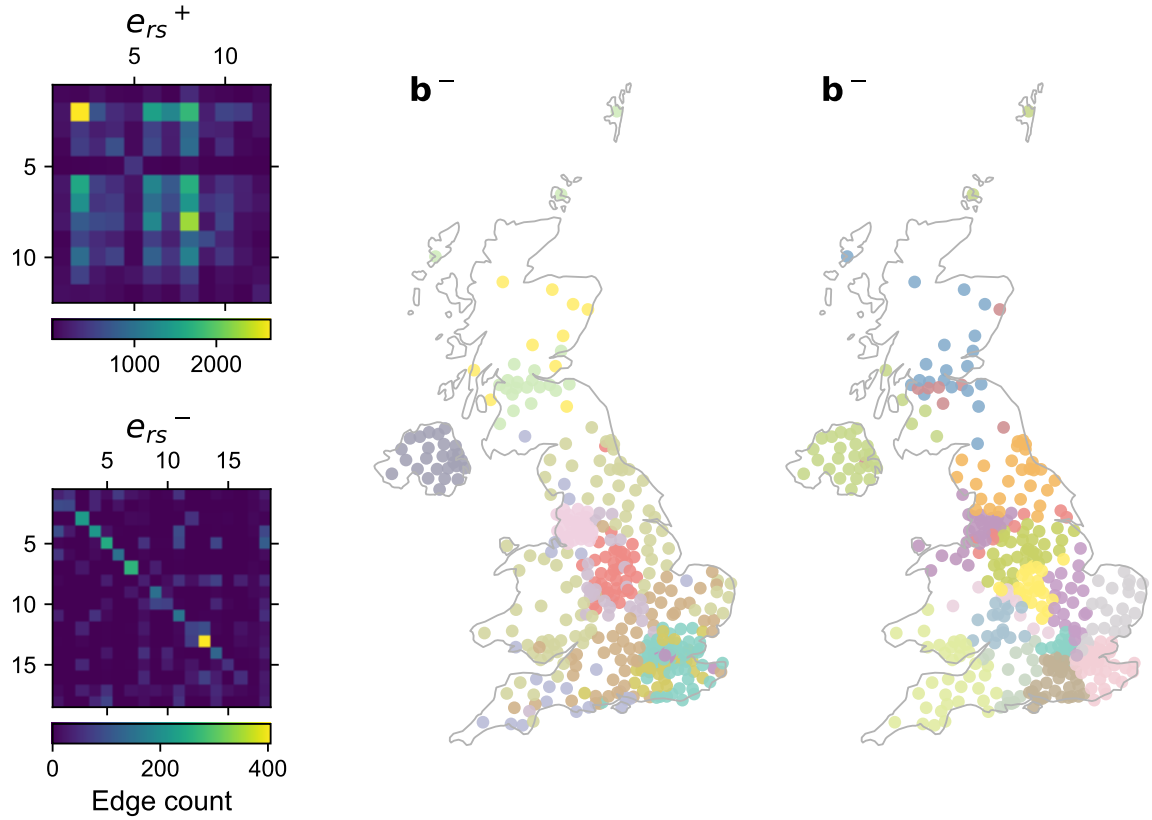


Figure 4.13: Minimum description length partitions and edge count matrices for the positive and negative backbones, as based on the doubly-constrained radiation model.

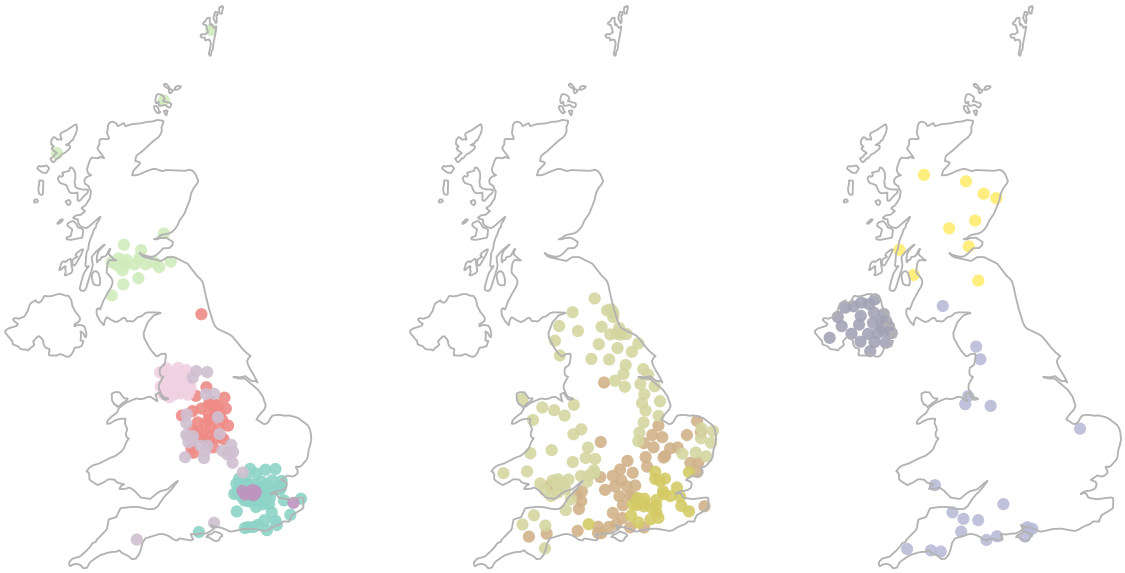


Figure 4.14: The positive partition $\mathbf{b}^+$ of the DC-radiation model is schematically broken up to facilitate distinguishing what the composition of the different blocks is.

blocks comprising Northern Ireland, the remote counties of Scotland, and a varied set of nodes in England and Wales close to the coast. This concentric organization seems to match well our intuition of the intervening opportunities and the analogy of the radiating particle under which the model was originally derived. As such, we find a good level of overlap between the peripheral communities, as measured using the Jaccard matrix.

We note in passing that the two brightest entries on the diagonal $\mathbf{b}^+$ correspond to blocks 2 and 8, respectively South-East London and Manchester. Furthermore, it is striking how England is perfectly separated from Scotland but for a single county in light blue that belongs to the outermost periphery.

We compute the average per capita GDP inside each block (in Figure 4.15), and as before we found the most interesting results for the negative partition. This time the block corresponding to inner London is an even larger outlier because the method groups Westminster and the City of London (per capita GDP of over £100 thousand) only with two other nodes (Camden, and Kensington and Chelsea).

Other blocks with a high average include the group of Belfast, Glasgow, Edinburgh and Aberdeen, the group containing Birmingham and other Midland counties, and the industrial North including Manchester, Liverpool and Leeds. Though there are some differences node for node, we recognise these groups because we also found them in the partition $\mathbf{b}^-$ which was based on the gravity model (in that case Liverpool was not placed with Manchester and Leeds, Sheffield was, and this is interesting because Sheffield and Leeds are roughly at the same distance from Manchester).

The partition $\mathbf{b}$ obtained from the layered SBM can be seen in Figure 4.16. It has $B = 17$ blocks and its similarity to the signed partitions is respectively $\mathrm{NMI}(\mathbf{b}, \mathbf{b}^+) = 0.604$ and

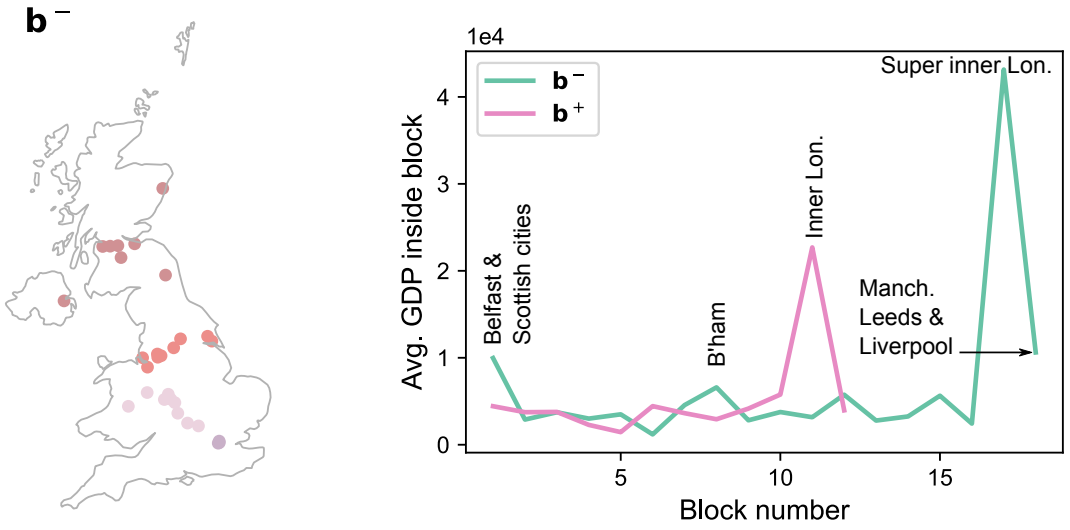


Figure 4.15: (Left) Highlighted blocks in the negative backbone. (Right) Average GDP computed for the nodes assigned to a given block.

$\mathrm{NMI}(\mathbf{b}, \mathbf{b}^-) = 0.758$, thus that the negative backbone has a bigger influence even though there are many more positive edges for the spatial backbone based on the radiation model. Indeed, we do not find as much the cores and the peripheries of the positive partition while at the same time the divisions between the countries are sharper than they are in $\mathbf{b}^-$.

Picking up on this point, we note that Wales has been almost perfectly separated from England and partitioned into two blocks (the single exception is Flintshire in North-East Wales that gets grouped with nearby Liverpool and Manchester). These two blocks are striking in that they pick up the cultural subdivision with the cities in the South (Swansea, Newport, Cardiff) being grouped together and completely surrounded by the less densely populated counties. The first group has a mean county population of 161 thousand and an average density of 7.5 residents per hectare, while the latter has respectively 114 thousand and 0.95. What is even more remarkable is that this signal is not present in $\mathbf{b}^+$ nor $\mathbf{b}^-$, it only appears once we consider both signs of edges simultaneously with the layered SBM.

4.6 Application to the retail network

We previously saw the scatter plots with the observed and expected number of trips (for the doubly-constrained models) inside Figure 3.6. The CPC and RMSE of the gravity model are 0.726 and 133, compared to 0.706 and 133 for the radiation model. With these values, it is no longer obvious that we should prefer a particular model because of a better fit.

Instead, we base our decision on a practical consideration: with the gravity model we have a significantly sparser backbone because 31% more edges are considered not statistically significant when using the exact tests (cf. Table D.1 in Appendix D). This is caused by the lower topological and weighted densities, respectively $\rho_t = 0.164$ and $\rho = 7.77$, which, according to our validation, are quite low for the gravity model and this has the counter-intuitively desirable consequence

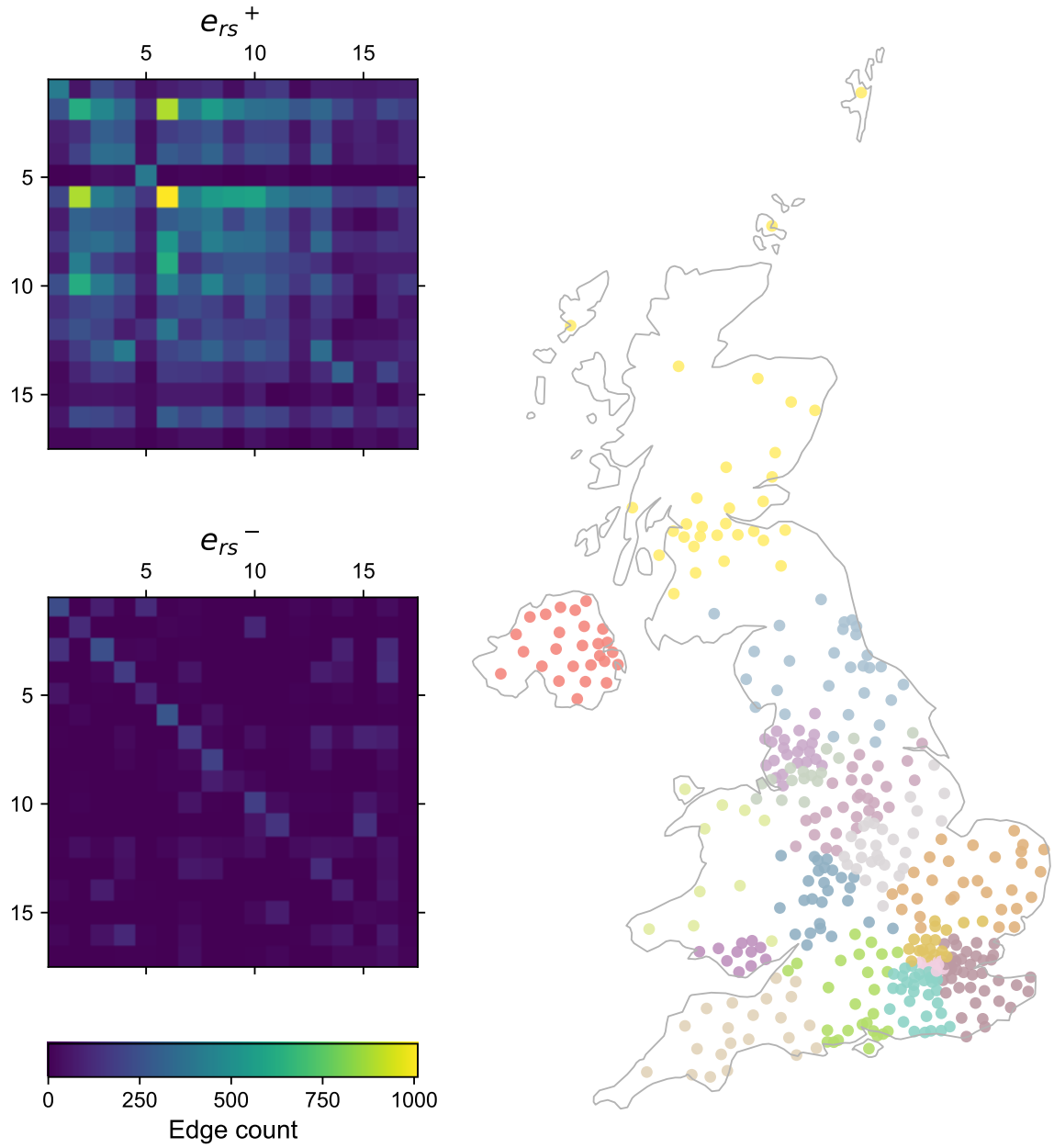


Figure 4.16: Minimum description length partitions and edge count matrices for the layered SBM, as based on the doubly-constrained radiation model.

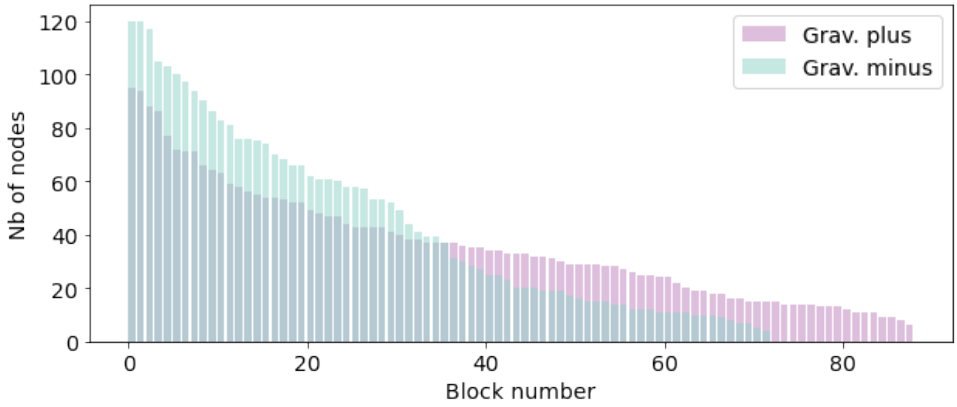


Figure 4.17: Block sizes of the positive and negative partitions for the doubly-constrained gravity-based backbone.

that that only the strongest signals remain. In the following section, we describe these results.

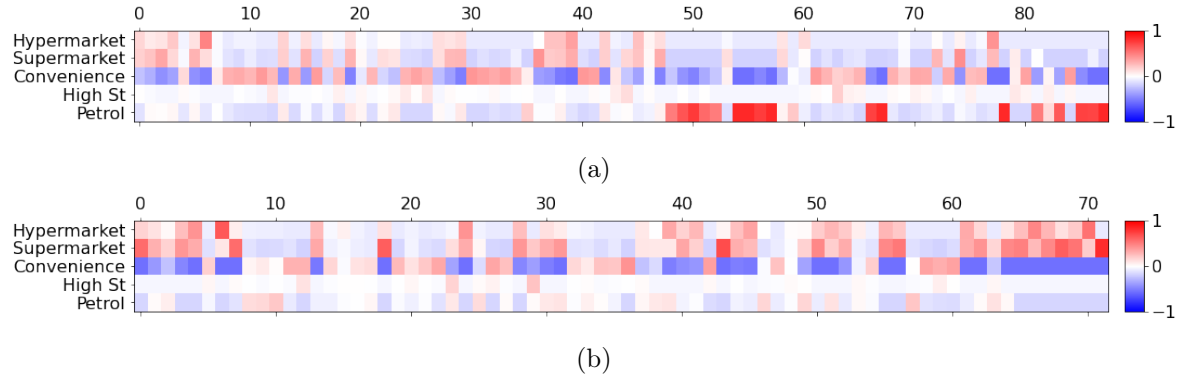
4.6.1 Gravity partitions

We apply our method to the spatial backbone based on the doubly-constrained gravity model. Like we did before, we fit the DC-SBM a hundred times and keep the partitions with the overall minimum entropy. The positive partition $\mathbf{b}^+$ has 88 blocks while $\mathbf{b}^-$ has 72 blocks. We find $\mathrm{NMI}(\mathbf{b}^+, \mathbf{b}^-) = 0.684$, indicating some level of agreement between the two partitions.

There are too many blocks to visually inspect them by colouring all the nodes at the same time and we inspect these partitions differently than for the commuting example. We show the number of nodes inside the blocks in Figure 4.17. The groups sizes are generally more balanced for the positive network.

The stores have different formats which roughly correspond to their size and product assortment (and indirectly with their position with respect to town). The largest formats are Hypermarket (257 stores across the country) and Supermarket (545). The next format by size is High Street which is also the least common (104 stores), and the smallest formats are Petrol (493 stations) and Convenience which is by far the most common format (1,754 stores).

We compute the format composition of the blocks in each partition, normalise them as proportions summing to one, and we take the difference with the fractions for the whole country. The results are presented in Figures 4.18a and 4.18b. In both figures, we note there are very few instances where the first two rows have opposing colours inside the same column. This means that the largest formats, Hypermarket and Supermarket, show a correlated behaviour and they are disproportionately more abundant or lacking inside the blocks almost always at the same time. This finding makes sense because it is unlikely that customers distinguish between the two formats and it is more likely that they use one or the other indistinguishably. Another interesting signal that we can see in Fig. 4.18a are the brightly coloured red entries in the bottom row. These are about a dozen blocks in the positive backbone that are composed of two thirds or more of petrol stations, and these are located all over Great Britain (see Fig. 4.19c). Intuitively, it makes sense that our method can detect groups of refuelling shops because these



(a)

(b)

Figure 4.18: Comparison of the block composition in terms of store formats when compared against the average for the UK: (a) for the positive backbone and (b) for the negative backbone.

are not likely to be visited one after the other.

Using the Jaccard index of the convex hulls of the blocks in $\mathbf{b}^+$, we locate two regions that are separated into groups that have a high degree of overlap (there are many more of these and there are also some in the negative backbone). The first region includes the cities of Manchester, Liverpool and Lancaster (excluding Sheffield and Leeds) and the second region surrounds Norfolk. Both regions are broken up into three large blocks which we highlight in Figure 4.19b. The sub-matrix of edge counts for these blocks can be seen in Figure 4.19a, and Table 4.1 contains the difference with average composition for the UK for the six blocks of interest (these values are plotted as columns inside the matrix in Fig. 4.18a).

Though both regions have a different number of nodes, they have a remarkably similar mesoscale structure. Using the numbers in Table 4.1, we can see that in each region we find one block that has a larger share of Hypermarket and Supermarket stores, one block with a larger share of Convenience stores, and one block that is almost entirely made of Petrol stations. Since we are dealing with the positive backbone, the entries of the $\mathbf{e}^+$ matrix (Fig. 4.19a) indicate that there are many more trips than expected directed to — and originating from — the groups of petrol stations in each region. There is also a disassortative structure when we focus on blocks number 30 and 55 which are the Convenience stores and Petrol stations in Norfolk, with almost no links inside each group but a significant number of links from Convenience to Petrol and the other way around. Finally, we see small but non-vanishing values in the entries that link the large formats in one region to the Petrol stations in the other region (e.g 13 to 55 and 28 to 66), a signal that we can probably associate with travelling customers.

Another interesting block that we were able to find in the positive backbone is block number 75, composed of 11 nodes. This is the community highlighted in Figure 4.19d. Most of the nodes are located in Edinburgh but the remaining ones are at the Glasgow Airport, just outside Belfast Airport or in Clapham, London. This grouping of nodes very much reminds us the one we found with the Scottish cities and Belfast for the commuting network.

We would like to reflect on the importance of the methodology presented here and to do so, we ask the reader to remember the composition of the communities found using the map equation (seen in Section 2.3.5). In our first attempt to describe the structure of the customer-

(a) (b) (c) (d)

Figure 4.19: Highlighted blocks of the positive partition $\mathbf{b}^+$. The figures (a) and (b) should be taken together: in (a) we have a submatrix of edge counts corresponding to the six blocks highlighted in (b) and that were selected because of their high overlap and mesoscale organisation as discussed in the text (the colours above the matrix serve as a legend). We also highlight in (c) 13 blocks that contain a disproportionately high number of petrol stations and in (d) a single block which is composed of stores in Scotland, Belfast and London.

	Block nb	Block size	Hyper.	Super.	Convenience	High St	Petrol
Manchester,	13	55	**0.173**	**0.191**	-0.411	-0.033	0.080
Liverpool and	14	95	-0.60	-0.110	**0.328**	-0.012	-0.146
Lancaster	66	29	-0.082	0.173	-0.453	-0.033	**0.740**
Norfolk	28	38	**0.103**	**0.275**	-0.372	0.020	-0.025
	30	32	-0.081	-0.173	**0.412**	-0.002	-0.156
	55	13	-0.082	-0.173	-0.556	0.033	**0.844**

Table 4.1: Difference between the format composition for the block highlighted in Fig. 4.19 and the average composition for the United Kingdom. We have highlighted interesting values, as discussed in the text.

store network, we found 31 groups of nodes that were clustered geographically. Now, we have seen that there are different regions — Norfolk on the one hand, and Manchester and Liverpool on the other — that are organised in a similar way, with three blocks that have primarily different formats and which are visited with some type of avoidance or affinity for particular successive transitions between formats (the most obvious of which is that a Petrol station is less likely to follow a visit to a different Petrol station).

This is only one example, but what we have hopefully proved here is the potential for the space-corrected SBM to uncover signals beyond the purely geographical. Our wish is is now for Dunnhumby, who has a greater visibility of the *Retailer*'s business as well as other sources of data, to build upon our work and extend the study of how the loyal customers shop across the network of stores.

Chapter 5

Spatio-temporal analysis of shopping habits

So far, we have studied the retail dataset from the perspective of where the shopping is taking place and what are the interactions between the stores. In the previous two chapters, we employed a projection of the originally bipartite customer-store network and we kept only the store nodes. Now, we would like to go back to studying the shopping behaviour of individual customers.

The anonymised shopping records contain a great deal of untapped information about the way in which people shop, and this holds true even if we do not look at the specific products bought. By focusing, for example, on the location and the time of shopping, we are able to reconstruct an approximation of the movement of the customers when they are out buying for groceries.

This chapter started as an exploration of the mobility of customers, and it evolved into a study of various measures that can be calculated at a basket level. We are interested in spatial and temporal patterns related to how customers normally shop, and how the COVID-19 pandemic made them change their habits. The pandemic has led to many studies that focus on human mobility using GPS or call records [41, 62, 105, 109], and it showed scientists working in complex systems that more than ever they have a role to play informing policy-makers [15, 56, 79], so this is our modest contribution with a hopefully complementary source of data.

5.1 Definition of the measures

The anonymised shopping records take the form of a large table which we represent mathematically as $\mathcal{B}$, the set of all transactions. An element $\mathfrak{b} \in \mathcal{B}$ — also called a basket — is an ordered list which holds a timestamp, the *household id* that identifies the customer, the *store id*, the *store format*, the *channel* (offline or online), the *mission*,[1] the amount paid which we call *spend*, and the *number of items* bought. The timestamp and the two ids (household and store) uniquely identify a basket.

[1] Missions are a proprietary type of basket segmentations that we will encounter in more detail below.

For notational convenience, we define functions that extract the relevant information of the ordered list such as the household ID, $\text{hid}(\mathcal{b})$, or the store ID, $\text{sid}(\mathcal{b})$, etc. (The names are self-explanatory.)

As explained elsewhere, online baskets are still served from physical stores[2] and they appear in our dataset mixed with offline baskets. We distinguish the two using the binary variable $\text{channel}(\mathcal{b}) \in \{\text{Online}, \text{Offline}\}$. We also define

$$\mathcal{F} = \{\mathcal{b} \in \mathcal{B} : \text{channel}(\mathcal{b}) = \text{Offline}\}, \tag{5.1}$$

$$\mathcal{O} = \mathcal{B} \setminus \mathcal{F}, \tag{5.2}$$

respectively the set of in-store and online baskets.

Obviously, for any measure to accurately reflect the habits of customers, we need to observe a large enough portion of all the shopping done. We can expect that what we can attribute to each household using the loyalty card is in reality just a fraction of all their shopping because there will be times when they do not use their card or they visit a competitor.

This is where our definition of loyal customers comes in handy, because the households that are in our dataset have been shopping regularly at the *Retailer* for the better part of four years. With this in mind, we have some degree of confidence that we are observing the majority of the shopping (even if we are not able to quantify how much), otherwise the customers would not be classified as loyal.

5.1.1 General metrics

We now introduce a number of measures that characterise the volume of shopping and the use of channels, formats, etc. Given h, the ID of a selected household, we denote their set of baskets as

$$\mathcal{B}_h = \{\mathcal{b} \in \mathcal{B} : \text{hid}(\mathcal{b}) = h\}. \tag{5.3}$$

The number of visits is the cardinality of this set,

$$n(h) = \left|\mathcal{B}_h\right|. \tag{5.4}$$

Moreover, the number of online and offline shops are, respectively,

$$n_o(h) = \left|\{\mathcal{b} \in \mathcal{O} : \text{hid}(\mathcal{b}) = h\}\right|, \tag{5.5}$$

$$= \left|\mathcal{B}_h \cap \mathcal{O}\right|, \tag{5.6}$$

and

$$n_f(h) = \left|\{\mathcal{b} \in \mathcal{F} : \text{hid}(\mathcal{b}) = h\}\right|, \tag{5.7}$$

$$= \left|\mathcal{B}_h \cap \mathcal{F}\right|. \tag{5.8}$$

[2]To be precise, online baskets are only served from stores that belong to the two larger formats: Hypermarket or Supermarket.

Obviously, we have that $n(h) = n_f(h) + n_o(h)$ so only two of these quantities are independent. We find it useful to calculate the proportion of online shopping done by household h as

$$p_o(h) = n_o(h)/n(h)\,. \tag{5.9}$$

Turning our attention to the stores in which household h buys,

$$S(h) = \{s \in \mathcal{S} : \mathrm{sid}(\mathcal{b}) = s,\ \ \mathrm{hid}(\mathcal{b}) = h\}\,, \tag{5.10}$$

the number of different locations visited is

$$n_S(h) = \big|S(h)\big|\,. \tag{5.11}$$

The average time that a household waits between successive shops is

$$\Delta(h) = \frac{1}{n}\sum_{k=2}^{n} |t_k - t_{k-1}|\,, \tag{5.12}$$

and, related to temporal information, the household proportion of weekend shops is

$$p_w(h) = \frac{1}{n}\big|\{\mathcal{b} \in \mathcal{B} : \mathrm{hid}(\mathcal{b}) = h,\ \ \mathrm{day}(\mathcal{b}) \in \{\mathrm{Saturday}, \mathrm{Sunday}\}\}\big|\,. \tag{5.13}$$

A household's average basket spend is calculated using

$$\langle \mathrm{spend}\rangle(h) = \frac{1}{n(h)}\sum_{\mathcal{b}\in\mathcal{B}_h} \mathrm{spend}(\mathcal{b})\,, \tag{5.14}$$

and their average number of items shopped per basket is

$$\langle \mathrm{items}\rangle(h) = \frac{1}{n(h)}\sum_{\mathcal{b}\in\mathcal{B}_h} \mathrm{items}(\mathcal{b})\,. \tag{5.15}$$

Finally, a household's average spend per item is given by

$$\langle q\rangle(h) = \frac{1}{n(h)}\sum_{\mathcal{b}\in\mathcal{B}_h} \mathrm{spend}(\mathcal{b})/\,\mathrm{items}(\mathcal{b})\,. \tag{5.16}$$

We also define four other metrics that are entropy measures. We remember that the entropy of a discrete random variable X is calculated as

$$H(X) = -\sum_{x} \mathbb{P}(X = x)\log\mathbb{P}(X = x)\,. \tag{5.17}$$

For household h, we view the day of the week when they shop as a random variable D_h, taking values over the set $\{\mathrm{Monday}, \mathrm{Tuesday}, \ldots, \mathrm{Sunday}\}$. Using Equation (5.17), we calculate the *day-of-the-week entropy* as $H(D_h)$ with the probabilities equal to the normalised frequencies of each day.

We also define the random variable F_h for the format of the shops visited, taking values over $\{\text{Hypermarket}, \text{Supermarket}, \text{Convenience}, \text{High St.}, \text{Petrol}\}$ and the *format entropy* is $H(F_h)$. The *store entropy* is $H(S_h)$, where $S_h = S(h)$ is the set of stores defined in Equation (5.10).

Baskets are segmented (i.e. categorised) by the *Retailer* using so-called missions. The are 6 of them: *Replenish*, *Few days*, *Full shop*, *Food now*, *Food later* as well as the special (but not very useful) category *Undetermined*. The segmentation has to do with the type and number of products inside a basket. Even though we do not know exactly how missions are calculated, for the sake of our analysis it will be enough to consider that the first three missions refer to assorted baskets that differ in size (a small basket to top-up on few items, a medium-sized basket that should last 2 or 3 days or a full shop that should cover roughly a week) and the other two refer to food that is bought to eat either now (think a sandwich bought during lunch) or later (think a prepared meal that will be eaten for dinner but bought during the lunch break). We represent the mission using the random variable M_h and the *mission entropy* is obtained as $H(M_h)$.

Store metrics: Some of the metrics above can be also be applied to stores. Let s denote the ID of a particular store and let $\mathcal{B}_s = \{\mathscr{b} \in \mathcal{B} : \text{sid}(\mathscr{b}) = s\}$ be the set of baskets shopped by all the customers at s. In the same way we had before, the number of visits is $n(s) = |\mathcal{B}_s|$.

We find it useful to distinguish the total number of physical visits to s and the number of online orders served from s, respectively

$$n_f(s) = \left|\mathcal{B}_s \cap \mathcal{F}\right|, \tag{5.18}$$

$$n_o(s) = \left|\mathcal{B}_s \cap \mathcal{O}\right|. \tag{5.19}$$

The total spend (which we do not find it useful to calculate for households) is given by

$$\sigma(s) = \sum_{\mathscr{b} \in \mathcal{B}} \text{spend}(\mathscr{b}), \tag{5.20}$$

and the average basket spend is

$$\langle \text{spend} \rangle(s) = \frac{1}{n(s)} \sum_{\mathscr{b} \in \mathcal{B}} \text{spend}(\mathscr{b}). \tag{5.21}$$

We note that we do not control for inflation which, over the period 2017-2021, we estimate to be at an annual average ranging from 1.5% to 2.5%. While this is definitely a shortcoming that we need to keep in mind, we believe that it does not affect in a crucial way our analysis because it is based on describing general trends.

5.1.2 Mobility metrics for in-store shopping

We now introduce the measures which quantify the travelling of household h and that are computed using only the offline shopping records.

For a particular transaction $\mathscr{b} \in \mathcal{B}_h \cap \mathcal{F}$ that was made in-store, we can cross-reference its store ID with a list of coordinates and locate household h in space and time using $\boldsymbol{\rho}(\mathscr{b}) =$

$(x, y, t) \in \mathbb{R}^2 \times \mathbb{R}^+$. We then order all the transactions in $\mathcal{B}_h \cap \mathcal{F}$ using their times $t_1 < \ldots < t_{n_f}$, and view them as stops. With the positions indexed as $\boldsymbol{\rho}_k = (x_k, y_k, t_k)$, we define the trajectory of household h as $\boldsymbol{\gamma} = (\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_{n_f})$.

We also find it useful to represent the geographical coordinates by $\mathbf{r}_k = (x_k, y_k) \in \mathbb{R}^2$, where the sub-indices indicate that the corresponding timestamp of the basket is t_k.

While the above framework works for general trajectories, we note that we can only locate households by placing them at stores, and each $\mathbf{r}$ actually references the coordinates of a store. In Section 5.3, we use $||\mathbf{r}_i - \mathbf{r}_j||$ as shorthand for distance separating stores i and j (indexed, for example, using the same ordering we had for the OD matrix in Chapters 3 and 4).

The *centre of mass* is the average position of the household over the period of interest.

$$\mathbf{r}_{\mathrm{CM}}(h) = \frac{1}{n_f(h)} \sum_{k=1}^{n_f} \mathbf{r}_k \,. \tag{5.22}$$

The *radius of gyration* represents the linear size occupied by a trajectory [39] and it is calculated as the square root of the variance of the position,

$$r_{\mathrm{G}}(h) = \left(\frac{1}{n_f(h)} \sum_{k=1}^{n_f} ||\mathbf{r}_k - \mathbf{r}_{\mathrm{CM}}||^2 \right)^{1/2} \,. \tag{5.23}$$

The (linear) *distance travelled* is the discrete line integral

$$d(h) = \sum_{k=2}^{n_f} ||\mathbf{r}_k - \mathbf{r}_{k-1}|| \,. \tag{5.24}$$

Let us now say a word about the distances. We have mentioned in passing that the coordinate associated to a selected transaction is a point $\mathbf{r} \in \mathbb{R}^2$, and this might seem odd because we were previously calculating the great-circle distance based on the longitude and latitude of the stores. However, in the customer-centric calculation of the radius of gyration, the centre of mass, etc., we only need to calculate the distance between stores that a given household visits, and these are typically only a few kilometres apart.

We find it easier to project the coordinates and calculate Euclidean distances. We therefore convert latitude and longitude in the World Geodetic System of 1984 (EPSG:4326) to Easting and Northing of the United Kingdom Ordnance Survey (EPSG:27700) with units in metres. This time we do not do anything special to the stores that have overlapping coordinates.

5.1.3 Restriction to specific time windows

The metrics previously introduced make use of all of the shopping records made in four years by a single household and summarise them into a single scalar number. However, in order to study what is their evolution over time, we need to restrict ourselves to the shopping that occurred within specific time periods.

For household h and time window $[t_s, t_e)$, let

$$\mathcal{B}_h(t_s, t_e) = \{\mathcal{b} \in \mathcal{B} : \text{hid}(\mathcal{b}) = h,\ \ t_s \leq t(\mathcal{b}) < t_e\}\,, \tag{5.25}$$

denote the subset of baskets happening in the period of interest. Most of the metrics generalise in the obvious way, making use of the baskets in $\mathcal{B}_h(t_s, t_e)$ instead of $\mathcal{B}_h$. For example, $n(h; t_s, t_e) = |\mathcal{B}_h(t_s, t_e)|$; $n_o(h; t_s, t_e) = |\mathcal{B}_h(t_s, t_e) \cap \mathcal{O}|$; and so on and so forth.

For the centre of mass and the radius of gyration, Equations (5.22) and (5.23), we employ the set of indices

$$\mathcal{K} = \{k \in \mathbb{N} : t_s \leq t_k < t_e\} \tag{5.26}$$

and we take the sums over $k \in \mathcal{K}$ and also replace $n(h)$ by $n(h; t_s, t_e)$.

Lastly, for the distance travelled, Equation (5.24), suppose that the times whose indices are in the set $\mathcal{K}$ can be ordered as $t_u < \ldots < t_v$. The distance travelled restricted to the time window is

$$d(h; t_s, t_e) = \sum_{k=u+1}^{v} ||\mathbf{r}_k - \mathbf{r}_{k-1}||\,. \tag{5.27}$$

To choose the time periods, the most important consideration is that the library Pandas in Python provides a fast and powerful way to group the records whose timestamp occurs within a periods that line up with the calendar month. Therefore, it is very convenient to group all of the shopping records that occur every calendar month (t_s and t_e are the first day of every month) or every calendar half-month (t_s and t_e are respectively the first and the fifteenth days of every month, or vice-versa).

This does mean that the time-windows have unequal lengths. Table 5.1 summarises which metrics are intensive and which extensive, and for the latter we normalise them by dividing by the number of days in the corresponding time window.

5.1.4 Containing geometry: centroid tree algorithm

Given a point $\mathbf{x} \in \mathbb{R}^2$ and a division of the country into D administrative geometries[3] we have applications that will need us to determine which geometry contains $\mathbf{x}$. The excellent `shapely` library in Python makes it easy to determine this calling the `contains` method on each of the geometries and passing the point $\mathbf{x}$ as an argument.

The complexity of this search when we have H points is $O(HD)$. The issue that we face is that H can be an extremely large number (think several hundreds of thousand when we are dealing with all the households). We propose instead the following algorithm that takes as input the centroids, the point $\mathbf{x}$ and a parameter k:

1. Compute the centroids of the D divisions.
2. Construct 2-d tree with the centroids. This is a space-partitioning data structure that allows us to find the nearest neighbours amongst the centroids in logarithmic time.

[3] A geometry is any type of vector spatial data. Here we take it to mean a Polygon or Multi-Polygon object in `shapely` whose vertices have coordinates in the reference system that we are using.

	Metric name	Symbol	Intensive/Extensive
Households	Nb. visits	$n(h)$	E
	Nb. online shops	$n_o(h)$	E
	Prop. online	$p_o(h)$	I
	Nb. unique stores	$n_S(h)$	I
	Time delta	$\Delta(h)$	I
	Prop. weekend	$p_w(h)$	I
	Avg. spend	$\langle \text{spend} \rangle(h)$	I
	Avg. items	$\langle \text{items} \rangle(h)$	I
	Avg. spend per item	$\langle q \rangle(h)$	I
	Day-of-week entropy	$H(D)$	I
	Format entropy	$H(F)$	I
	Store entropy	$H(S)$	I
	Mission entropy	$H(M)$	I
	Radius of gyration	$r_{\text{G}}(h)$	E
	Tot. distance	$d(h)$	E
Stores	Nb. visits	$n(s)$	E
	Nb. physical visits	$n_f(s)$	E
	Nb. online shops	$n_o(s)$	E
	Total spend	$\sigma(s)$	E
	Avg. spend	$\langle \text{spend} \rangle(s)$	I

Table 5.1: The different household and store metrics, and whether they are intensive or extensive variables.

3. For each point $\mathbf{x}$, use the tree to return the k nearest centroids.

4. Use the corresponding k geometries as candidates and loop over them to test whether $\mathbf{x}$ is contained inside any of them.

The complexity of this algorithm is $O(Hk \log D)$ which in practice can decrease the running time of computations from a few days to a few hours. The cost that we pay is that we are not guaranteed to find the containing geometry amongst the k nearest centroids. However, with $k = 12$ the worse failure rate that we encountered was 1.4% for the month of April 2020 (we could not determine the containing geometry for 8,339 households amongst more than 580 thousand).

5.2 Changes in the mobility and shopping behaviours due to the COVID-19 pandemic

Equipped with the metrics presented in Section 5.1, we now study their evolution in time.

The shopping records that we had access to are organised by fiscal years, covering 2017 to 2020, and when we refer to a year we mean fiscal and not calendar year. The *Retailer*'s fiscal year roughly starts at the end of February but the date changes each year. Because of this, we append the last few days of February (the beginning or fiscal year n) to the records of the previous year (the year $n-1$), so as to normalise the four years and have them start in March.

By a happy coincidence — if there can be one — this comes in handy when we focus on the changes due to the COVID-19 pandemic. Indeed, its effects were first felt in the UK during

March 2020 so that fiscal years line up with the start of the pandemic. For reference, we now give a timeline of some of the most relevant restrictions that were in place in England during the first year of the pandemic.[4]

1. First national lockdown (26 of March to 1st of June 2020)
2. Lower restrictions (1st of June to 4 of July 2020)
3. Minimal restrictions (4 of July to 14 of September 2020)
4. Restrictions reintroduced (14 of September to 5 of November 2020)
5. Second national lockdown and then Tier 4 (5 of November and 19 of December 2020)
6. **Partially covered**: Third national lockdown (6 of January to 6 of March 2021)
7. **Not covered**: Leaving lockdown (March to July 2021)

5.2.1 Average trends for the customers

We begin our analysis looking at the average behaviour of the customers over the observation period of four years; we look at the seasonal variability of the metrics introduced in Section 5.1, averaged for the 586,416 loyal households in our dataset which are spread in all parts of the UK.

In Figure 5.1, we can observe some of the trends related to the visits and the money spent. We also have the two travel metrics in Figure 5.1 (compare these visualisations with the travelling in Denmark [4, 5]) and the entropy measures in Figure 5.3. Each solid line indicates the mean across all the customers for a year; the (fiscal) years 2017 to 2019 are in blue and the year 2020 in orange. The different tones of grey in the background indicate the severity of the lockdown restrictions put in place (described above), with darker shades indicating harsher stay-at-home measures.

General description Focusing on the blue curves in Figure 5.1, we find that the number of different stores peaks in August, and this is associated with a small dip in the number of visits during that same period, indicating that the summer months have a very special behaviour. For two of the three years, August is also the time when the average spend per item peaks.

There is a more dramatic fall in the number of visits during the first half of January. This follows the period of highly increased spending during the second half of December (and this is in-store spending, as seen from the fall in the proportion of online shopping), so we are seeing the aftermath of the Christmas holidays which are the most important period of celebration for a majority of the population.

Elsewhere we will talk about the correlation of the metrics in the context of clustering customers, but at this point we just note in passing that the curves for the average time Δ and the curves for the average basket spend look very much alike, and they also look like a reflection

[4] Adapted from the timeline published by the House of Commons Library: https://commonslibrary.parliament.uk/research-briefings/cbp-9068/

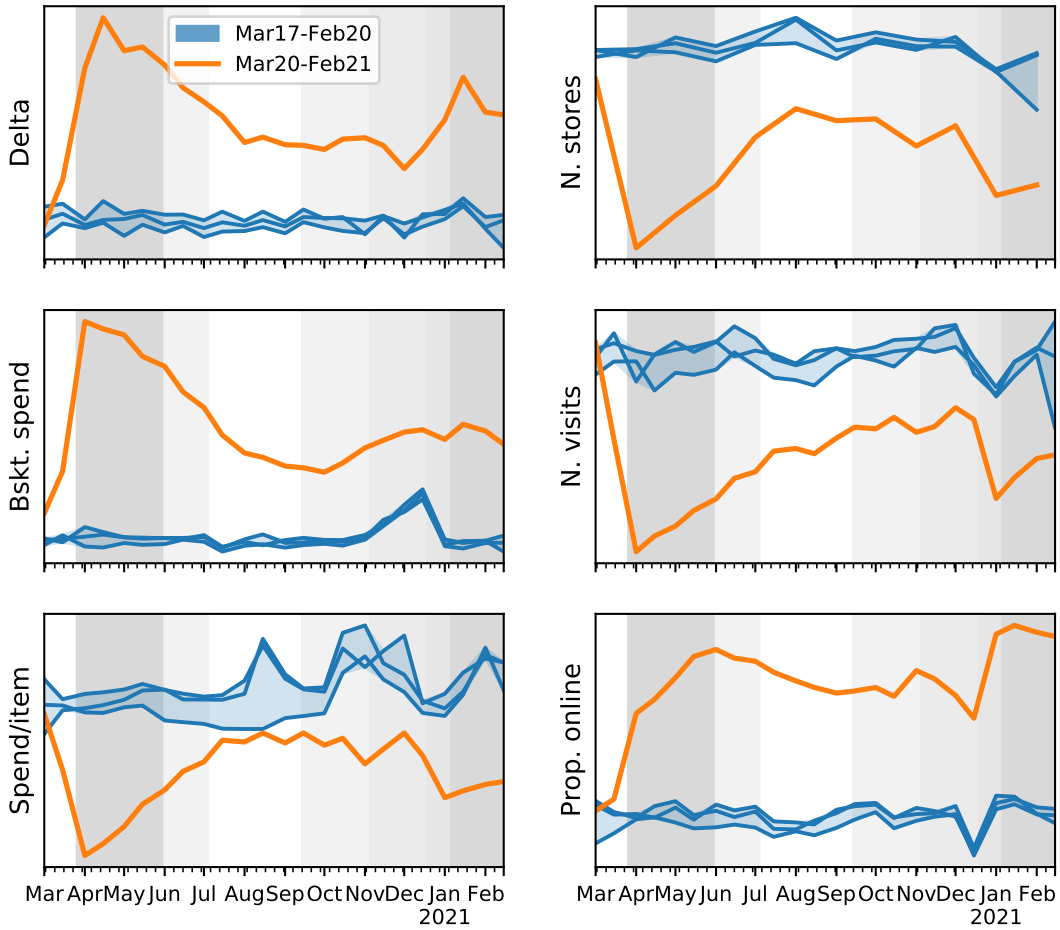


Figure 5.1: Time series for the visits-related metrics averaged for the whole of the UK (over 580 thousand loyal households).

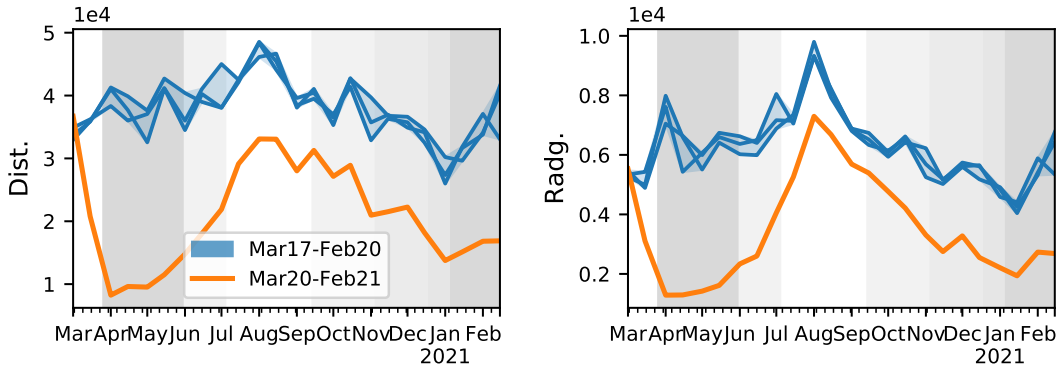


Figure 5.2: Time series for the travel-related metrics averaged for the whole of the UK (over 580 thousand loyal households).

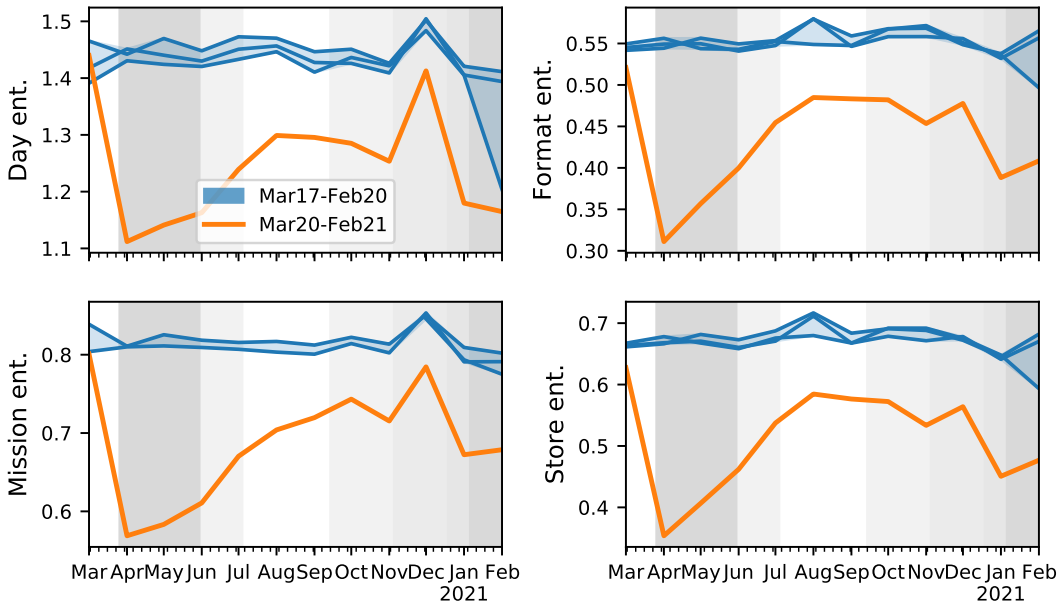


Figure 5.3: Time series for the entropy metrics averaged for the whole of the UK (over 580 thousand loyal households). The significant dip during the second half of February 2017 seems to be caused by having fewer records from 4 years ago (it is a fact that they are progressively deleted but it is hard to quantify how it affects us).

of the number of visits. This is a sanity check for us, because it makes intuitive sense that the larger a basket is, the more time it takes that household to go back to the shops and as a result the average number of visits drops.

In Figure 5.2, we concern ourselves with the mobility of households as measured using the linear distance (left) and the radius of gyration (right). We can see large peaks during the first half of April and over August, together with the overall nadir the first half of January. Again, the holidays have an important effect in the seasonal patterns which is altogether unsurprising.

For the entropy measures in Figure 5.3, we can see two general patterns, one for the day-of-the-week and mission entropies on the left side, and a second one for the format and store entropies on the right side. Though we do not know how the missions are calculated, intuitively it makes sense that shops in the same types of stores at given days are assigned the same mission. On the other hand, the number of formats visited goes hand in hand with the number of different stores visited, and one cannot increase without the other increasing. So overall there is some redundancy in these metrics (which again, we will cover in more detail when we do clustering).

For the day-of-the-week and mission entropies, the profile of the blue curves is mostly flat throughout the year except for a peak in December, meaning that the shopping in less predictable in terms of what it is that the customers are looking to buy (and which day of the week) during the Christmas holidays. For the format and store entropies, the peak occurs in the summer with a nadir in December. This is related to the behaviours we have seen already of an overall larger travelling done in August.

Name	Format	Country	Affluence level
University Town (HSt)	High Street	England	Mid
London Mid (S)	Supermarket	England	Highest
London Outer (H)	Hypermarket	England	Highest
London Outer (P)	Petrol	England	NA
Manchester Inner (C)	Convenience	England	Mid
Regional (S)	Supermarket	Scotland	Highest

Table 5.2: The list of six stores that we look at in detail and some of their characteristics. The affluence level is not defined for the Petrol format.

Changes due to the pandemic It is clear from every panel in the three figures seen so far that (12 metrics in all) that 2020 is unlike anything seen the previous years: the size of the baskets increases dramatically, a larger proportion of the shops are done online and both of these things go hand in hand with a smaller number of visits to fewer stores. All of the entropy measures decrease indicating more predictable shopping patterns, while the mobility decreases two or three-fold due to the national restrictions.

Roughly speaking, these patterns follow the amount of restrictions that were in place from March 2020 to February 2021. The peaks in spend and the nadirs in visits coincide with the lockdowns put in place at the end of March and at the end of December 2020; the number of different stores visited and the radius of gyration are closest to their pre-pandemic levels over the summer months when there are no restrictions. Interestingly, the number of visits continues climbing after the period of no restrictions as if there was a momentum in the recovery, but it seems that only this metric captures such behaviour.

We note that the proportion of online shopping increases in April but it does not peak during the first lockdown and it continues to climb until June. This is due to the massive logistic change that online shopping entails. Indeed, the *Retailer* had to increase its capacity and this took time so that in the beginning of the pandemic it was still very hard to get a delivery slot.

Overall, one of the most interesting aspects is that some of the changes mentioned above seem to be more permanent. While the radius of gyration during August 2020 is at a higher level than all the other months and it reaches levels that would be normal any other month besides August during the previous three years, it seems like the trend for the proportion of online shopping plateaus (with normal fluctuations like a decrease over Christmas) at a much higher level than before. Furthermore, even though we see some recovery happening over the summer months of 2020, almost all of the metrics are well separated from the previous three years. Considering that we are only describing the first year of the pandemic, it would be interesting to see how these changes are holding up more than two years after the pandemic hit the UK.

5.2.2 Trends for representative stores

We now take the viewpoint of the stores. Obviously, it is not practical to look at what happens to more than three thousand stores, so for the time being we only focus on a handful of stores that our industrial partner selected because of their location and format. These are the names listed in Table 5.2.

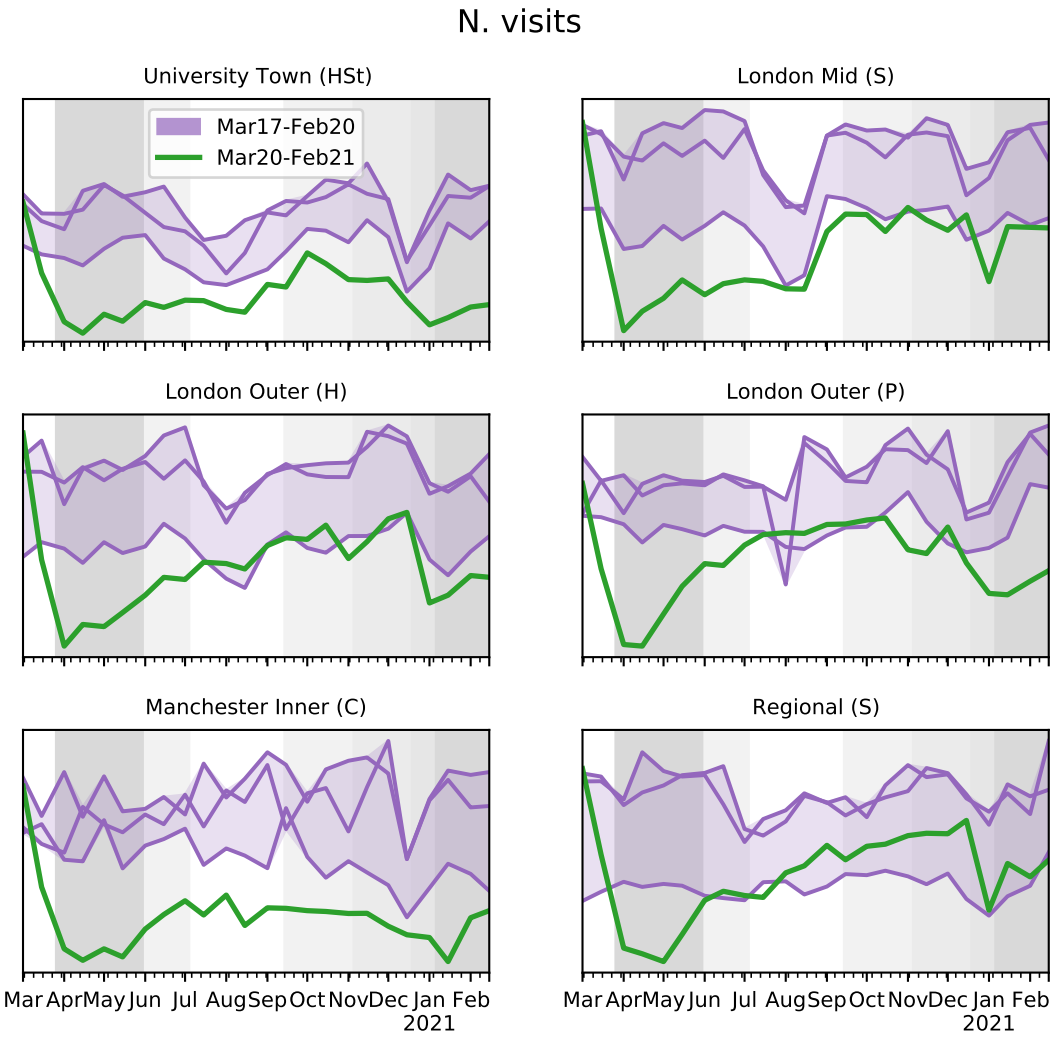


Figure 5.4: The yearly evolution of the number of visits for six selected stores. For anonymisation reasons, the values on the y axis are not shown.

The time series of the visits to our selection of stores can be found in Figure 5.4. One of the clearest patterns is the cyclical pattern of holidays in the pre-pandemic years, which is striking for the University Town and London Mid stores. The decrease in the number of visits in the University Town is largest in December, while the drop in London Mid is more pronounced over the summer.

In contrast, the convenience store Manchester Inner (located in the city centre) the drop that we can observe is over Christmas but the Easter and Summer breaks cannot be seen. The Regional shop located in a large city in Scotland has pretty constant demand all year round.

The Petrol station London Outer is located just outside the Hypermarket London Outer. Their curves are clearly correlated, though for the Petrol station the drop in July is not as significant as the one for the Hypermarket and there is also a peak in August that the Hypermarket does not have.

In terms of what happens after the pandemic starts, we compare how the green curve nears the purple ones for the different stores. The two stores that are located in the city centres, the High Street University Town store and the Convenience Manchester Inner never fully recover.

For London (Mid and Outter), there is some degree of recovery during the summer months and perhaps until November. Finally, the recovery is the most pronounced for the Supermarket (the largest format) Regional store in Scotland, with the number of visits increasing steadily from June until mid-December when the second lockdown is put in place. Furthermore, the drop due to this lockdown does not result in a drop that leaves the number of visits lower than the lowest year.

While it would be tempting to attribute this recovery to the COVID policy adopted in Scotland which was different to the one in England, we cannot dismiss the role of the format. The other large store among our list, the Hypermarket London Outer, also experiences an almost uninterrupted increase in the number of visits (actually starting earlier, in April 2020) until mid-December. This could be another of the more permanent changes induced by the pandemic: we know that household are shopping larger baskets in a reduced number of stores, so we hypothesize that the larger formats are the biggest winners during the pandemic year of 2020.

Compared to the number of visits, the total spend (the revenue from the loyal customers) in Figure 5.5 tells a different story. The pre-pandemic patterns are similar to what we have seen, with the largest difference being that the peaks in December are very pronounce for the London Outer and Mid stores.

In 2020 however, the High Street University Town store, the Supermarket London Mid and the Hypermarket London Outer never experience a decrease that puts them below the lowest level of the three previous years (the University town does during the second lockdown). In contrast the revenue at the Petrol station decreases very significantly presumably because the population is buying less petrol, and the Regional store actually sees and increase in demand. This means that format is an important factor to determine whether the revenue increases at a given store but the effect cannot be separated from the location of the (as seen from the fact that the store in the city centre of Manchester does experience a decrease in contrast to the

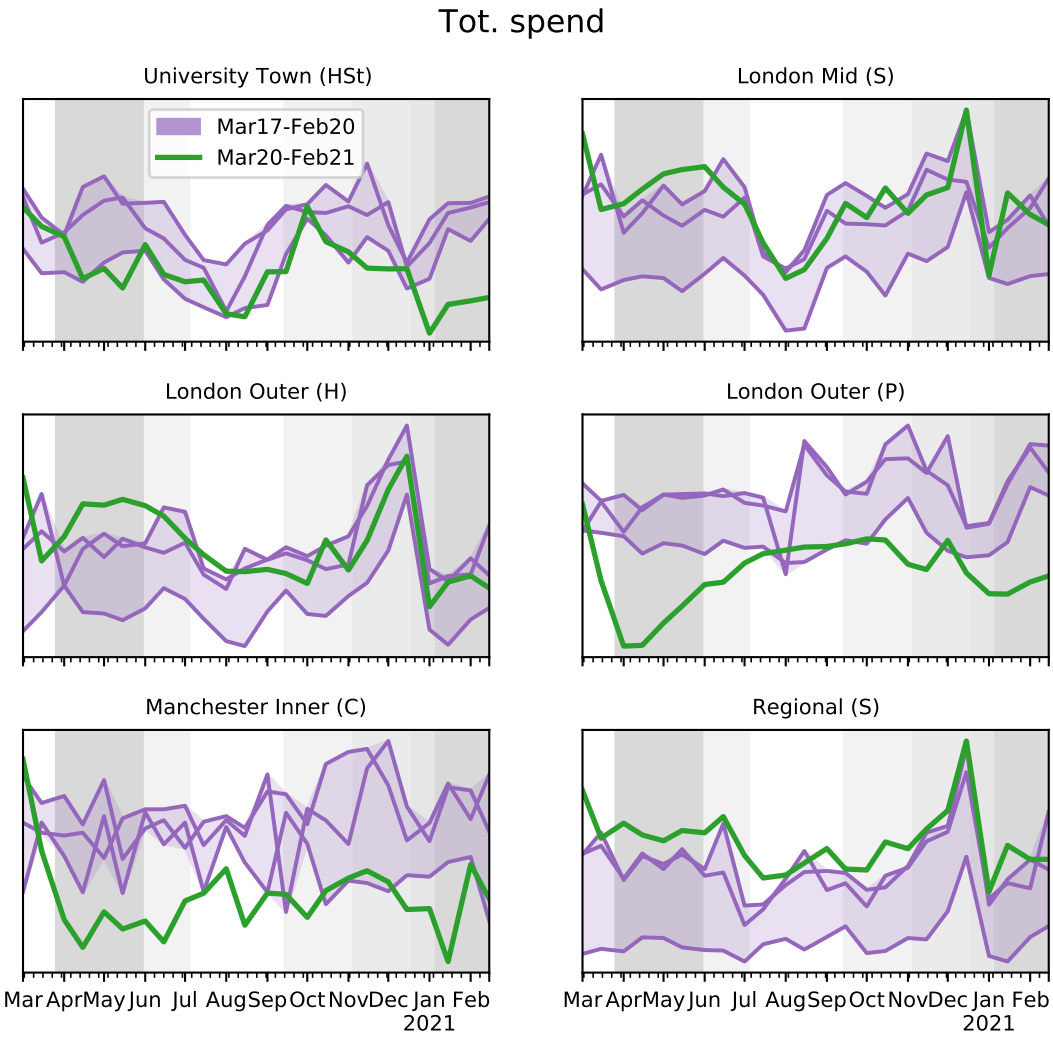


Figure 5.5: The yearly evolution of the total spend for six selected stores. For anonymisation reasons, the values on the y axis are not shown.

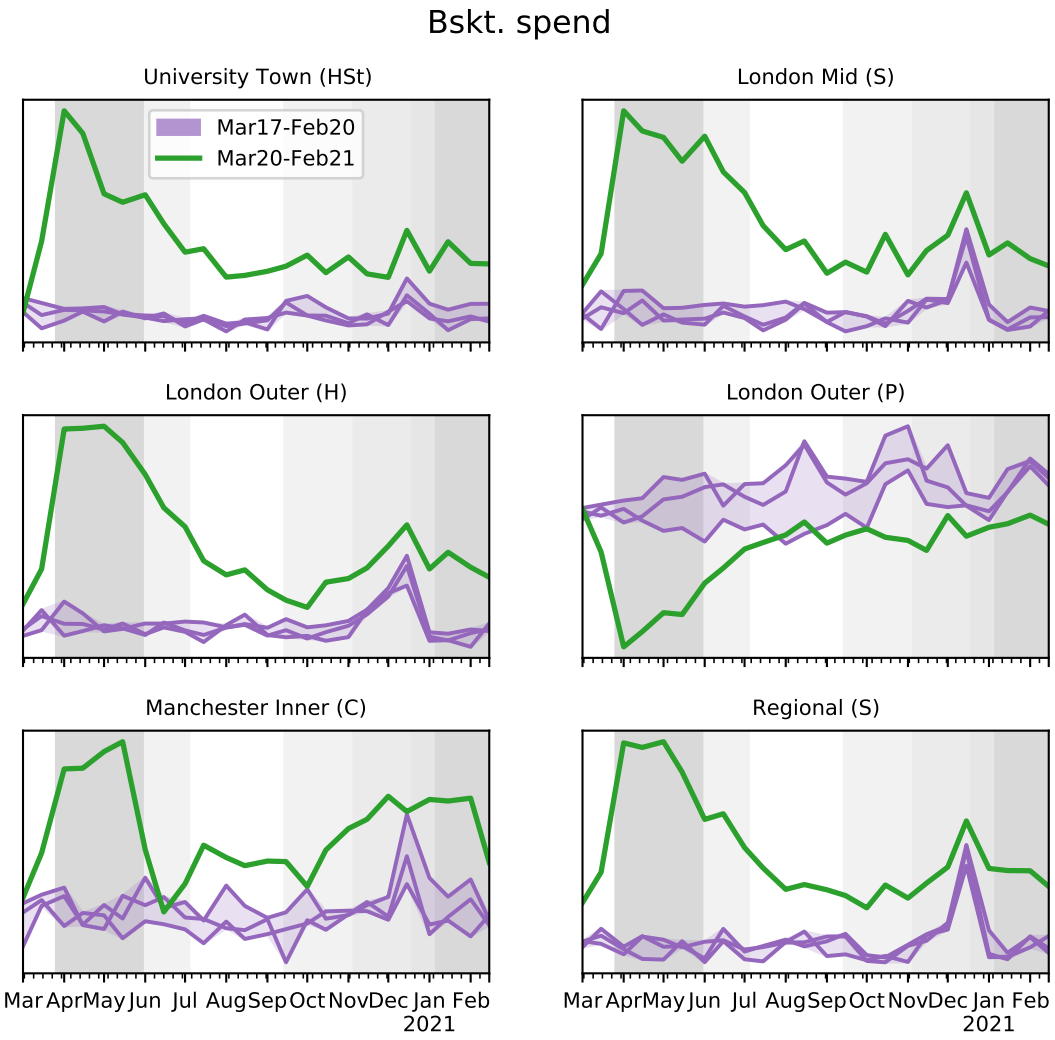


Figure 5.6: The yearly evolution of the average spend per basket for six selected stores. For anonymisation reasons, the values on the y axis are not shown.

University Town store).

In Figure 5.6, we have the final elements to tell the full story of the changes due to the pandemic. Because of the shocks to the number of visits and the increased — or at least not greatly decreased — spending, the average spend per basket increases in 2020 for five out of six stores. The exception is the petrol station, presumably because before the pandemic this format concentrates one of the most expensive types of baskets composed of a single item (petrol) worth £20-£60 pounds. The decrease in the metric in 2020 means that there are cheaper baskets being bought, and this could tell us that there were customers buying their groceries at the Petrol station or charging significantly less petrol than normally.

5.2.3 Spatial aggregation

Because it would be impossible to study what happens to over three thousand individual stores, we rely instead on aggregating the measures at the level of regions. We use the NHS Clinical

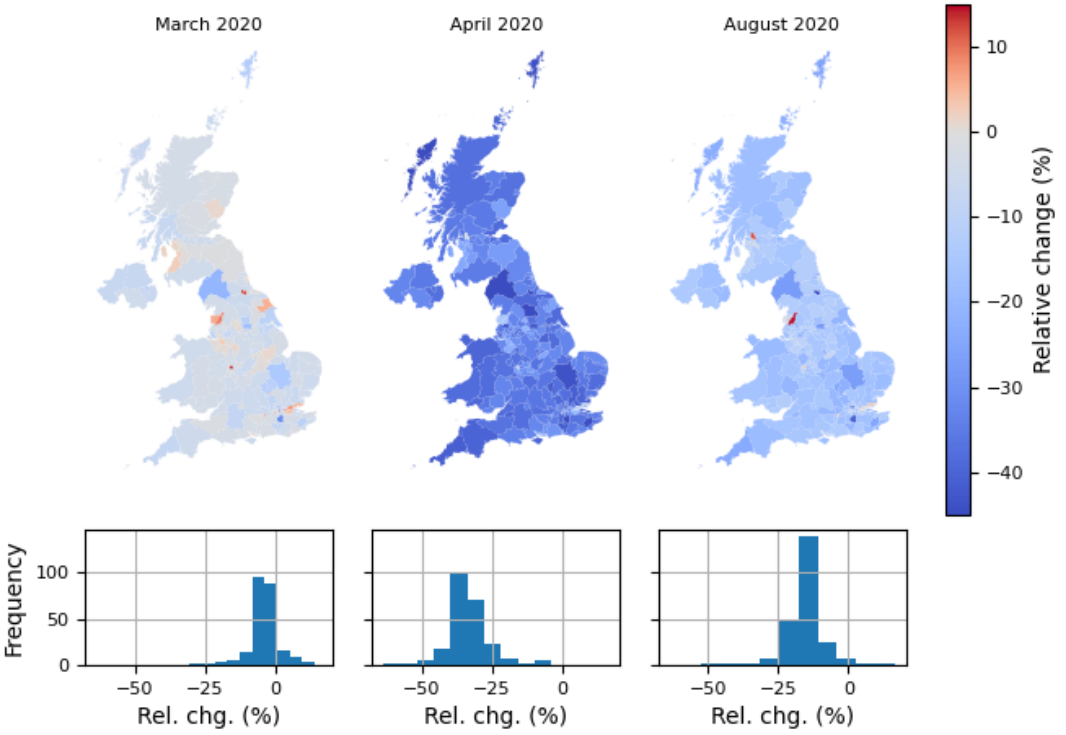


Figure 5.7: Regional change in the number of visits to physical stores during 2020.

Commissioning Groups,[5] also called Health Integration Authorities in Scotland,[6] Local Health Boards in Wales,[7] and Department of Health trusts in Northern Ireland.[8] These are the 234 administrative divisions of the NHS.

The main reason for selecting these administrative divisions instead of the local authorities (though the precise form of the geometries probably does not mask too much the spatial patterns) is that we were interested in doing a comparison with the Oxford COVID-19 mobility dashboard [9] which was tracking the average radius of gyration in each CCG calculated using GPS and mobile phone data [21]. Unfortunately the dashboard was taken down at the end of 2021 and we could not get any data, not even in processed form, from the researchers maintaining the dashboard.

The comparison with the GPS data would have been very interesting. The data source is significantly different to ours since the trajectories inferred from the supermarkets never really go back to a fixed location (home) like the GPS should. Furthermore, shopping for groceries would only be a fraction of all the mobility, though in the first lockdown in particular, it was probably one of the most important reasons for leaving home. We hope that in the future the COVID-19 dashboard is put up again, and a comparison with our data is still possible.

[5]The vector boundaries are freely available at https://geoportal.statistics.gov.uk/datasets/ons::clinical-commissioning-groups-april-2019-boundaries-en-bgc-1/about. The use of this data does not imply that we have any affiliation or support from the Greater London Authority.

[6]Boundaries freely available at https://data.gov.uk/dataset/c652c3b4-dcd3-4c9d-8738-e0934afb8953/health-integration-authorities.

[7]Boundaries freely available at https://data.gov.uk/dataset/b088818b-815f-42fd-bd45-95864aa3a71d/local-health-boards-april-2019-boundaries-wa-bfe.

[8]Boundaries freely available at https://www.opendatani.gov.uk/dataset/department-of-health-trust-boundaries.

[9]Previously available at https://www.oxford-covid-19.com/.

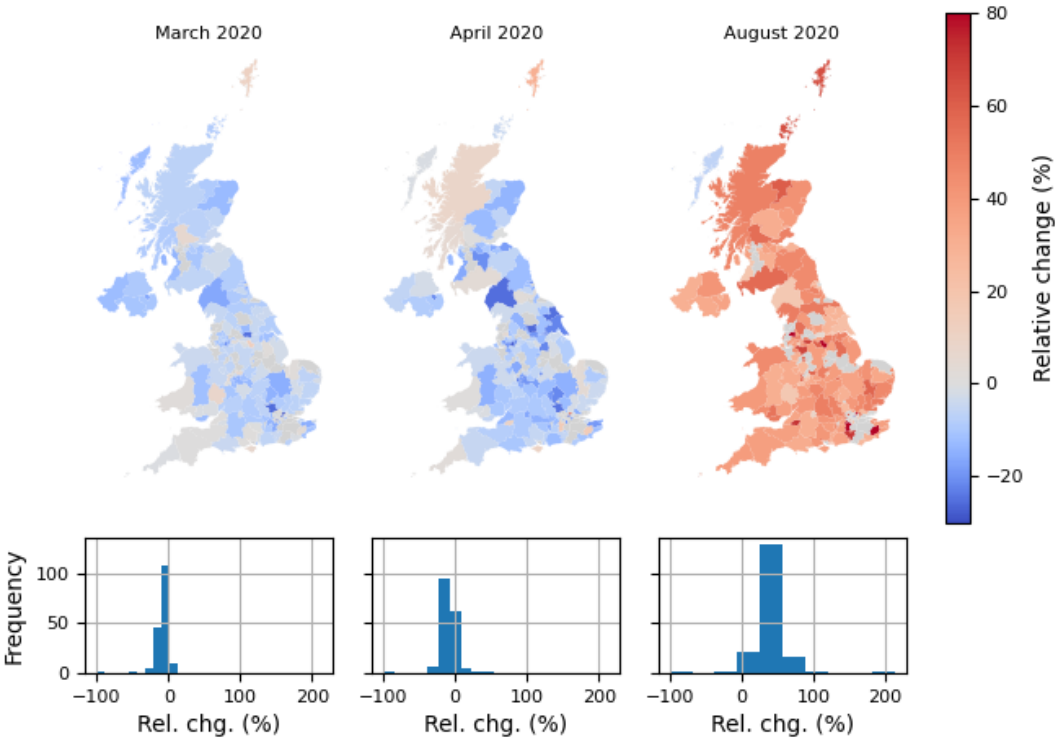


Figure 5.8: Regional change in the number of online shops during 2020.

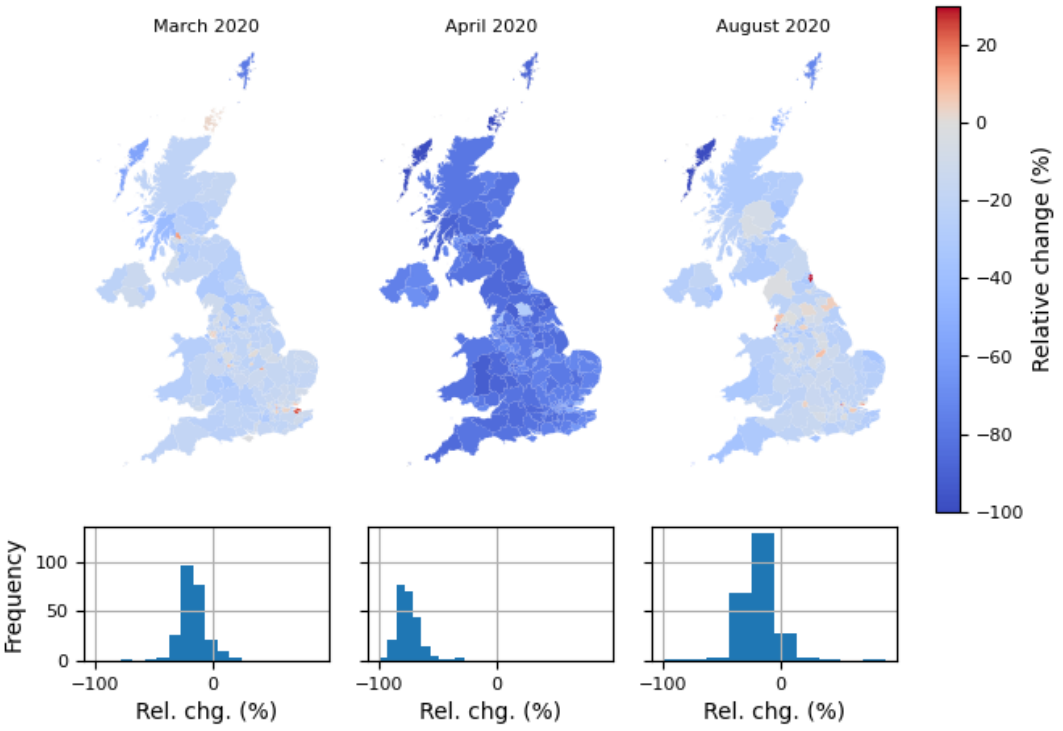


Figure 5.9: Regional change in the average radius of gyration during 2020.

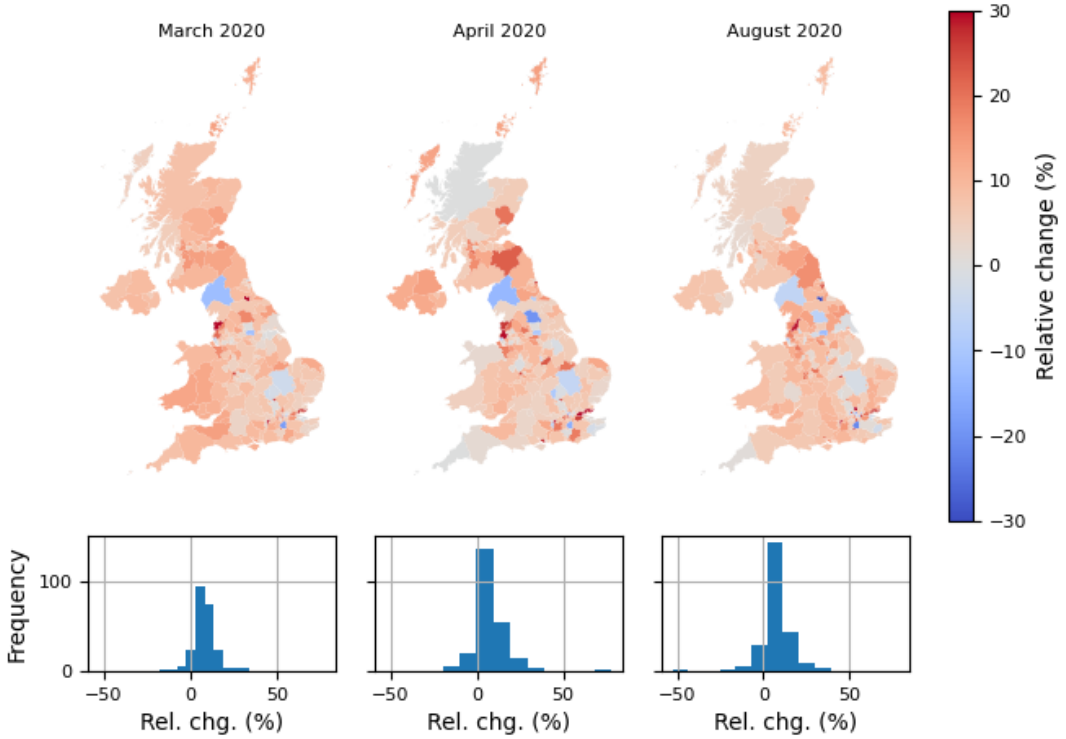


Figure 5.10: Regional change in the total revenue during 2020.

We compute the change due to the pandemic for four different metrics. For a particular month of interest, we calculate the mean of the metric for the years 2017, 2018 and 2019 and we use this as a reference for comparing the metric's value during that same month in 2020. We compare the months of March, April and August.

The physical visits to the stores can be found in Figure 5.7. In March we can see what amounts essentially to random fluctuations centrered just below zero (we can see the effect of the end of March pulling the results down). During the month of April, by contrast, there is a very significant drop (mean decrease: -33.7%; std: 7.7%).

In August we can already see some recovery but much like we saw for the time series, most regions are below their pre-pandemic average (mean decrease: -15.6%; std: 7.1%). It is worth noting that six regions have positive changes, but some of these can be attributed to random fluctuations because only two regions have an increase larger that 3 percent: Greater Preston in Lancashire (+16.6%) and West Dunbartonshire (+12.1%), to the West of Glasgow. We do not know what are the reasons for the these increases.

The online shopping is shown in Figures 5.8. We can see that online shopping generally increases, but it does not do so immediately (mean in April: -9.3%; std: 11.9%). This agrees with the delay we have seen in the time series for the proportion of online shopping. By August however, almost all of the regions have increased rates (mean increase: +40.3%; std: 22.7%).[10]

[10]We must note nonetheless that we are lacking data for 63 regions (a bit more than a quarter of all the regions) because of the regions without online shopping either during the month of August 2020 or during the three previous years. We also see a decrease of 100% for the Tower Hamlets CCG in London (the other negative entry is the Western Isles of Scotland with -6.2), which means that we should take this map with some degree of caution because it is likely that we cannot measure the online shopping from the loyal customers alone.

The radius of gyration can be seen in Figure 5.9. The drop in the mobility inferred from physical visits to the stores is very significant in April (mean decrease: -75.5%; std: 9.9%). Some recovery is underway by August (mean decrease: -18.1%; std: 15.8%), though we have seen from the time series that the national average for the radius of gyration remains well below what would be expected during the summer months. Spatially, we can see that most of the country is still experiencing a decrease.

There are only 17 regions that have experienced an increase compared to the previous years (all of them in England). Of these, the ones that have a significant increase are: South Tyneside (increased by 89.7%), Camden (+47.0%), Sunderland (+29.4%), Southport and Formby (+26.8%), and Southend (+19.2%). All of these regions are located in close proximity of the coast and not far from the cities of Newcastle, Liverpool and London and so it is not altogether implausible that this is due to being important recreation and leisure locations.

Finally, the total spend for each region is shown in Figure 5.10. Interestingly, the three months look mostly similar indicating that the spend over a month can change from store to store and between formats, but at the larger level of regions the seasonal variability is not very significant. We also remark that there is a increase across the country even during the month of March. This is quite surprising and it could be due to the normal inflation of prices as opposed to an effect of the pandemic.

5.3 Flow reconstruction

While the pandemic created a shock that changed many shopping patterns and the number of visits to the shops generally decreased, this shock was not homogeneous in space. We see this in Figure 5.11, where we show a map of London[11] with the relative change both in the number of visits and the spend during the month of April 2020, compared to the average for the same month the previous three years. Sune Lehman's group at DTU has visualised beautifully the change in the population landscape in Denmark [57] and we are striving to do something similar with the shopping data.

It is clear that central London and in particular Westminster, the City of London and their surroundings on the North Bank suffer the most significant drop in demand. The twenty shops that suffer the largest decrease in the number of visits and in terms of the monthly spend during April 2020 are the left columns respectively in Tables 5.3 and 5.4. For the visits, we have 12 postcodes that are either EC or WC; for the spend we have 9, and the majority of the remaining postcodes have low numbers such as SW1, W3 or E1.

On the other hand, a good number of the shops outside Central London experience an increase in spend and a less steep decrease in the number of visits. The twenty shops that experience the largest increase are on the right columns in the same two tables and we cannot find a single EC or WC postcode among these stores.

[11] In the figures that follow we will not show the actual locations of the stores for anonymisation purposes: both the x and y coordinates of the stores have random additive noise drawn uniformly from the interval $[-600, 600]$ (the units are in metres). This is true for the visualisation but for any calculation that requires the distance we still employ the real coordinates.

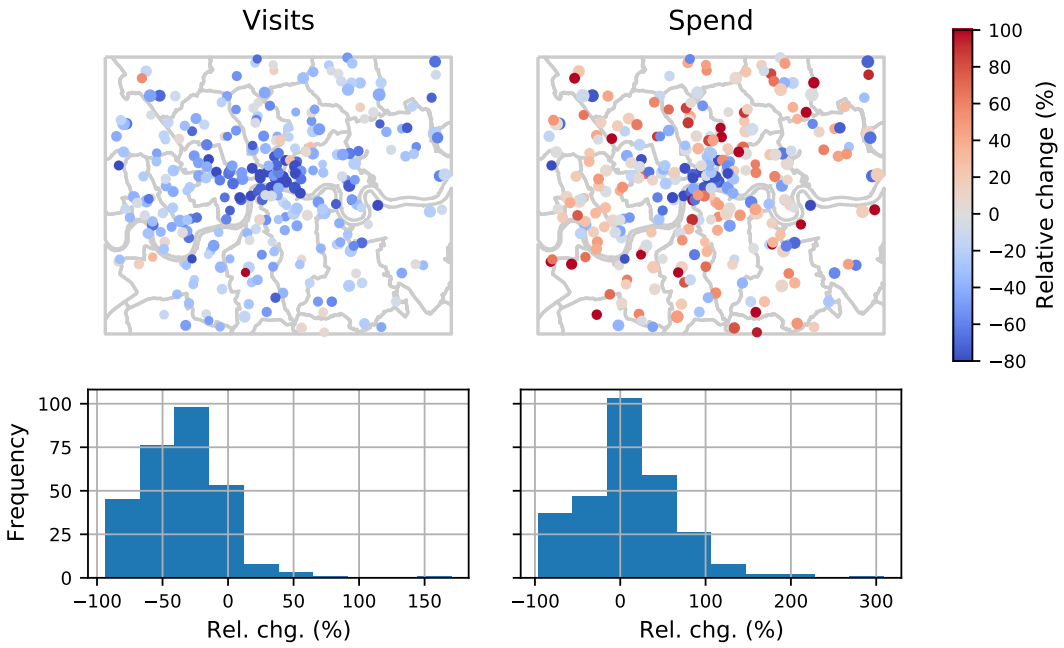


Figure 5.11: Relative change in the number of visits (left) and spend (right) during the month of April 2020, compared to the average for the same month during the previous three years. The coordinates have been changed uniformly at random inside a square of 1.2 km.

	Largest decrease			Largest increase		
	Format	Postcode	Rel. chg.	Format	Postcode	Rel. chg.
1	C	SW1	-93.7	C	SW2	170.4
2	C	SE10	-93.4	C	IG2	80.7
3	C	EC1	-92.6	C	NW4	55.8
4	C	W3	-92.0	C	CR4	54.5
5	C	SW1	-91.1	C	SW19	45.1
6	C	EC2	-90.7	C	SW15	33.1
7	C	E1	-89.0	C	E2	27.6
8	C	NW1	-88.9	C	E2	26.8
9	C	EC2	-88.5	C	N1	21.4
10	C	EC1	-87.6	C	SW15	17.1
11	C	WC2	-87.5	C	N8	16.5
12	HSt	EC2	-87.2	C	SE1	15.8
13	C	W1	-87.2	C	E10	15.1
14	C	EC1	-86.4	C	NW5	11.1
15	C	WC1	-86.1	C	SE11	9.7
16	HSt	EC3	-85.9	C	SE19	9.4
17	C	E1	-83.1	C	SW6	8.8
18	C	EC4	-81.8	C	SE26	8.4
19	C	SW18	-80.7	C	SE20	7.3
20	HSt	EC2	-78.8	C	SW14	6.9

Table 5.3: List of the shops in London that experienced the largest decrease and increase during April 2020 in terms of the number of visits.

	Largest decrease			Largest increase		
	Format	Postcode	Rel. chg.	Format	Postcode	Rel. chg.
1	C	SW1	-96.4	C	IG2	309.2
2	C	SW1	-89.7	C	SW2	215.8
3	C	EC1	-89.6	C	SW11	204.8
4	C	SE10	-87.7	C	SE10	175.1
5	C	W3	-85.4	C	SW14	171.8
6	C	E1	-85.1	C	E11	145.6
7	C	EC1	-83.0	C	SW19	141.5
8	C	EC2	-82.6	C	SE18	124.2
9	C	EC1	-79.5	C	SE4	122.2
10	C	WC1	-78.5	C	SE26	120.4
11	C	NW1	-77.3	C	NW4	120.4
12	C	SW18	-77.0	C	E11	114.6
13	C	EC2	-76.6	C	N16	106.8
14	P	NW2	-75.8	C	SE1	105.6
15	C	EC4	-75.6	C	SW15	104.3
16	HSt	EC2	-74.1	C	NW6	103.0
17	C	E7	-73.0	C	E2	102.8
18	P	IG6	-71.6	C	E8	101.0
19	C	WC2	-70.8	C	SE20	96.0
20	P	SE13	-70.0	C	IG2	93.2

Table 5.4: List of the shops in London that experienced the largest decrease and increase during April 2020 in terms of the total revenue.

We wonder if it is possible to establish a relationship between the stores that suffer a decrease in demand and the stores that see an increase. We can imagine that when the lockdown was put in place there were customers that shifted to working from home, and they probably changed some of the shopping they were making near their office or on their way there somewhere close to their home. Is there a way then to attribute that the visits or the revenue "lost" at the stores near their office was "gained" somewhere close to their home? In the population landscape changes seen for Denmark [57] it is intuitively clear that people are leaving Copenhagen and moving to smaller towns, but could this claim be formally supported with quantitative data?

Underlying our answer to these question — at least the one related to shopping — is the idea that it is plausible that the "shopping needs" of the customers remain approximately constant in time but that they are satisfied in different stores once the lockdown is put in place in April 2020 and many people stay at home. If true, we could observe the distribution of shopping made by a given household during two different time windows and infer if the distribution changed. In doing so, we can reconstruct our best guess of how the customers "moved" their shopping from a group of stores to other places.

The shopping needs are obviously hard to grasp, particularly without looking at the specific items, but in the analysis that follows we use either the number of visits or the spend by individual households as potential proxies for shopping needs.

While the total number of visits is unlikely to remain constant given the decrease we observe in Figure 5.11 (though online shopping cannot be seen in the map) and also the time series for

a selection of stores in Figure 5.4, the total spend might be a better proxy. Indeed, it is quite remarkable that the larger stores in Figure 5.5 like the Supermarket and the Hypermarket in London, or the High Street shop in the University Town, remain close to their pre-pandemic levels (and this is not the case for the number of visits) and we also saw in the choropleth map in Figure 5.10 that the three months look relatively the same in terms of the spend at the region level.

We will approach the flow reconstruction problem from the micro perspective of the empirical data recorded for individual customers, and then aggregate the results to obtain the macro view of the stores. The reason for this is that the algorithm that we use has a complexity of $\mathcal{O}(N^3)$ on the number of locations N, which makes it unfeasible to deal with thousands of stores at the same time (and typical customers only shop on a dozen of different locations at most).

5.3.1 Mathematical treatment

We employ the earth-movers distance (EMD) to reconstruct the flows that link the stores where visits or spend were lost to the shops where they were gained (we consider the subset of stores that a single household has interacted with). The EMD solves the following optimal transport problem: given two discrete distribution $\mathbf{p}$ and $\mathbf{q}$ (two vectors that have non-negative entries that sum to one) the *transport plan* is the solution to

$$\mathbf{F}^* = \underset{\mathbf{F}\in\mathbb{R}_+^{m\times n}}{\arg\min} \quad \sum_{i,j} f_{ij}\, c_{ij}\,, \tag{5.28a}$$

$$\text{s. t.} \qquad \mathbf{F}\mathbb{1} = \mathbf{p}\,, \tag{5.28b}$$

$$\mathbf{F}^\top\mathbb{1} = \mathbf{q}\,, \tag{5.28c}$$

$$f_{ij} \geq 0 \;\; \forall\, i, j\,, \tag{5.28d}$$

where $\mathbf{c} = (c_{ij}) \in \mathbb{R}_+^{m\times n}$ is the cost matrix, telling us how much it costs us to move a unit of mass from bin p_i to bin q_j.

The entry f_{ij} serves as a plausible reconstruction that moves the amount that we see is lost at location i to location j that gains mass. The hypothesis underlying this is that the reconstruction is more likely if it does not involve transferring the mass over long distances; in fact the most likely reconstruction has a minimal transport cost. This remains, without a doubt only an approximation of what really happened given our incomplete knowledge — in particular of the real location of the customers — but we believe that the minimization hypothesis gives us a sensible grounding to make the reconstruction problem well-posed.

Having seen the general formulation of the EMD problem, let us now explain our use for the shopping reconstruction problem. Suppose that we are interested in comparing the shopping done by household h during two different periods, $[t_s^{(a)}, t_e^{(a)})$ and $[t_s^{(b)}, t_e^{(b)})$. Let $m_a(i)$ denote the mass deposited at location i during the first time interval and similarly for $m_b(j)$. By mass we mean either the visits, the spend, or any other variable that we are studying that is distributed among the different stores; we just ask that our household has at least one location

with nonzero mass during at least one of the two time intervals (and this condition is always satisfied, otherwise we would not see the customer in our records).

Without loss of generality, we suppose that the total mass in the first interval is greater than the total mass in the second interval, and we write $\Delta m = \sum_k (m_a(k) - m_b(k)) \geq 0$.

We introduce the extended set of nodes $\mathcal{N}^* = \{0\} \cup \mathcal{N}$ where node 0 is a virtual node where will place the excess mass Δm. The rest of the nodes are ordered using their unique store ID, and we use the special ID 1 for all of online shopping (so we set $\text{sid}(\text{ɓ}) := 1$ whenever $\text{channel}(\text{ɓ}) = \text{Online}$); this ensures that the first node after the zero node corresponds to a second virtual node where online shopping takes place.

With the extended set of nodes, we define the distributions

$$p_a(0) = 0\,, \tag{5.29a}$$

$$p_a(i) = \frac{m_a(i)}{\sum_k m_a(k)} \quad \forall i \in \mathcal{N}\,, \tag{5.29b}$$

and

$$p_b(0) = \Delta m\,, \tag{5.30a}$$

$$p_b(i) = \frac{m_b(i)}{\Delta m + \sum_k m_b(k)} \quad \forall i \in \mathcal{N}\,. \tag{5.30b}$$

The distributions $\mathbf{p}_a$ and $\mathbf{p}_b$ respectively play the role of $\mathbf{p}$ and $\mathbf{q}$ inside Equation (5.28).

Next, we choose the cost function. We want node 0 to hold the mass as discarded from other locations so we let the cost of moving anything into node 0 be zero. We set the cost of moving the shopping between physical stores as the physical distance which is typically a few thousand metres. We did not experiment with alternatives, but its likely that other types of generalised distances — driving time, driving distance, a metric space embedding — would give sensible results. Moreover, it would be extremely interesting to compare the different answers.

For the cost of moving mass to the virtual node representing online shopping, we consider that it should be small because the decision to go online and shop for groceries (leaving aside the issue of getting a delivery slot which for the time being we do not model) introduces significantly less friction than having to shop at a new store which is a located a few kilometres away. We tested setting this cost to zero, but it resulted in degenerate reconstructions were only the online node sends or receives mass. However, as soon as we set the cost to a nonzero value which is less than the distances between the physical stores, we avoid this degenerate solution. We arbitrarily selected a cost of 1 unit (the units are in metres) but we tested that we would get the same results with a cost as high as 1000.

In summary, for stores $(i, j) \in \mathcal{N}^* \times \mathcal{N}^*$, the extended cost matrix is $\mathbf{c} = (c_{ij}) \in \mathbb{R}_+^{(N+1)\times(N+1)}$,

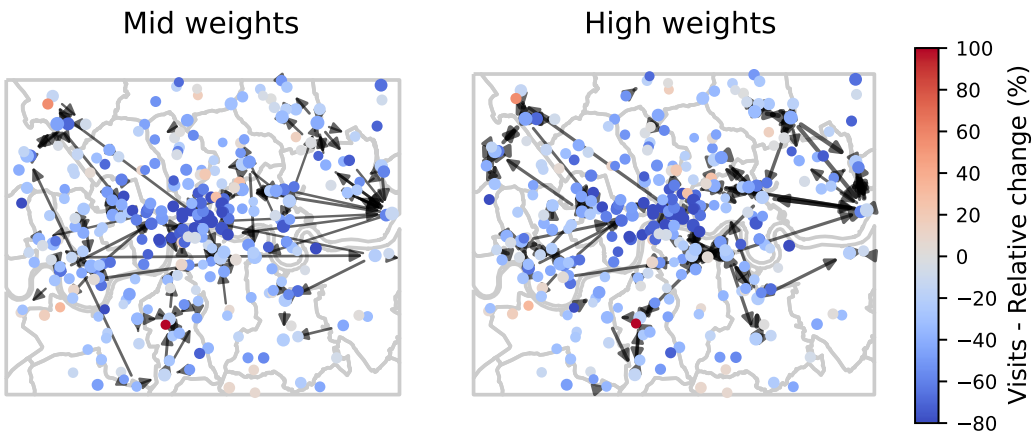


Figure 5.12: Reconstructed flow of demand in London from April 2019 to April 2020, measured using the number of visits. The colour scale of the stores represents the relative change we have seen before of April 2020 compared to the average of the three previous years.

with entries

$$c_{ij} = \begin{cases} 0 & \text{if } i = 0 \text{ or } j = 0\,, \\ 0 & \text{if } i = 1 \text{ and } j = 1\,, \\ 1 & \text{if } i = 1 \text{ XOR } j = 1\,, \\ ||\mathbf{r}_i - \mathbf{r}_j|| & \text{otherwise}\,. \end{cases} \tag{5.31}$$

With all the elements in place, we build for household h the distributions p_a and p_b, the cost matrix $\mathbf{c}$ and we solve (5.28). If for the selected household we have that $\Delta m < 0$, we reverse the roles of the two time windows and make sure at the end to take the transpose of the transport plan. As for the solution of the system, we employ the `ot.emd` routine provided by the Python Optimal Transport (POT) package [29] which implements the network simplex algorithm with complexity $\mathcal{O}(N^3)$. We are careful to implement a sparse version where the number of nodes is the union of the non-empty bins in the two time intervals and is never more than a dozen or so stores for the typical household. To deal with the whole population, we iterate over all the households in the dataset and then aggregate the reconstructed flows into a matrix that holds all the stores.

For simplicity and for consistency with the choropleth maps we examined before, we focus on the monthly frequency. The two time periods that we compare are the months of April 2019 and 2020 so that both time windows have the same length and they are exactly one year apart.

5.3.2 Reconstruction of the flows in London

We focus initially in London. We use the whole population to compute the matrix for the UK and then restrict ourselves to the flows that originate and terminate within the 293 stores in a rectangle that encompasses London.

The reconstructed flows obtained from the visits can be seen in Figure 5.12 and those from the spend in Figure 5.13. In both cases, we divided the flows into two panels so that the patterns are clearer, with the medium weights on the left and the higher weights on the right.

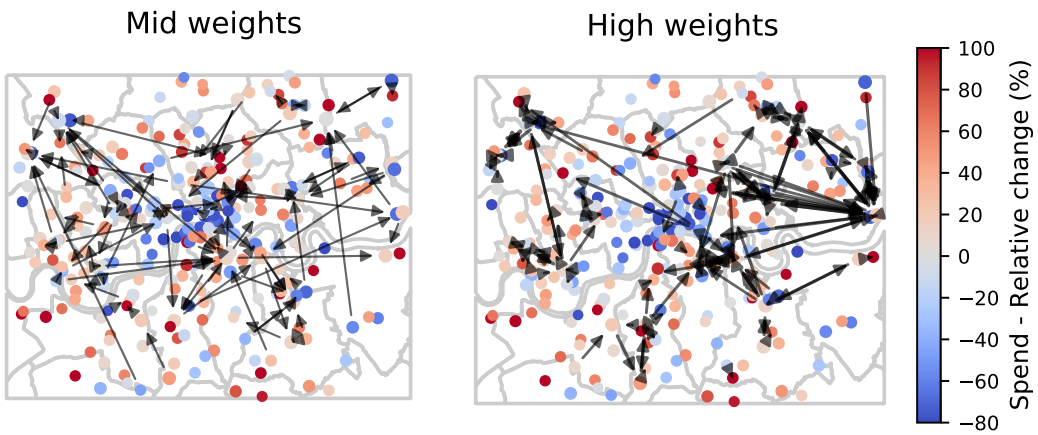


Figure 5.13: Reconstructed flow of demand in London from April 2019 to April 2020, measured using the total spend. The stores are coloured according to the relative change (April 2020 compared to the average three years).

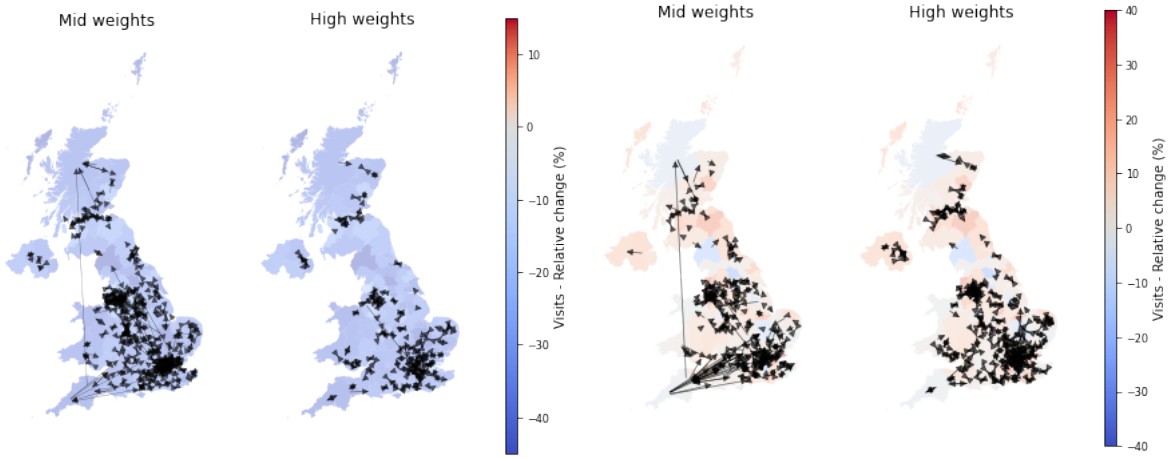


Figure 5.14: Reconstructed flow of demand between the CCGs measured using visits (left) and spend (right) from April 2019 to April 2020.

Amongst the mid-weights of the reconstructed flows of spend, we can see a large number of arrows whose origin is located around central London and that point outwards to other areas of the city, particularly the East or the North-West. The arrows with high weights for spend are generally shorter, with some exceptions that cross London from East to North-West. The height weights associated with short arrows intuitively make sense, because we expect that whenever possible the EMD will result in flows in close proximity due to the minimisation underlying the calculation.

For the visits (Fig. 5.12), it is harder to make out the pattern of outward pointing arrows from central London. Still, the North-West and the East seem to be the biggest receivers of flows like we had for the spend.

5.3.3 Aggregate flows at the CCG level

Next, we focus on the CCGs. We built the flows between the 234 regions by aggregating all the stores that are contained within each region. We continue to compare April 2019 and 2020; the flows due to the visits and spend can be seen inside Figure 5.14.

This time, the flows due to the visits look more like the flows due to the spend, possibly indicating that at the level of region it is indeed possible to measure the shopping needs either with visits or with spend. For the mid-weight flows, we can see longer arrows.

One of the most striking patterns is the bunch of arrows that originate in Cornwall and pointing all over South-East England. To make sense of these arrows, we remember that the reconstructed flows represent the shopping lost during the first period and gained somewhere else during the second time interval. Thus, we are seeing the shopping done in Cornwall during the Easter break in 2019 and "lost" to the shops near the home of the customers, because we know that a majority of the population was staying home in April 2020.

For the CCGs, it is even clearer than for London that the reconstructed flows with the highest weights lead to very short arrows and we see that most of the transfer of demand happened in close proximity.

5.4 Clustering of customers

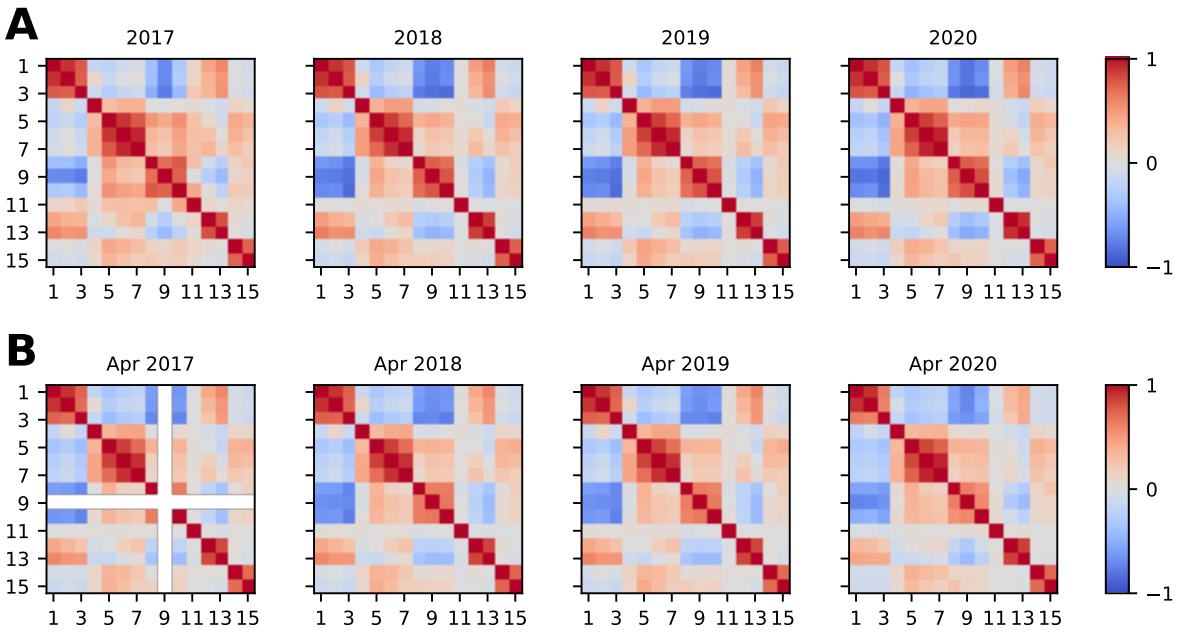
Both our industrial partner Dunnhumby and the *Retailer* rely on household segmentations to tailor specific promotions and cater to their customers' experience. While the segmentations are proprietary, we have been told that they are mostly based on the content and the price of the baskets and that currently there is nothing that that makes use or quantifies the mobility of the customers.

The segmentation of households, in a language more familiar to network scientists, is the problem of unsupervised learning of groups and we will now use the metrics defined above to run an exploratory analysis of household behaviour with clustering techniques. For ease of computation, we select a representative sample of households which amounts to one percent of the households.

5.4.1 Correlation of the measures

We have already discussed when analysing the time series that the fifteen metrics introduced contain redundant information and they seem to be correlated. In Figure 5.5, we represent the average correlation between the metrics for each year (Panel A) and also the month of April (Panel B) which we tend to focus on.

We discern groups of metrics that are highly correlated (we arranged the metrics so that this becomes more evident). In the first group, the average spend per basket is highly correlated with the average number of items bought (yearly mean correl.: 0.900; Apr. mean correl.: 0.902) and with the average time elapsed between successive visits to the stores (yearly: 0.783; Apr.: 0.711).



Legend

1. Avg. items	4. Avg. spend/item	7. Format ent.	10. DoW ent.	13. Prop. online
2. Avg. spend	5. Nb. stores	8. Nb. visits	11. Prop. weekend	14. R. of gyration
3. Time delta	6. Store entropy	9. Mission ent.	12. Nb. online	15. Tot. distance

Table 5.5: **A**. Yearly correlation matrices; the different dates (months for the calculation) are simply considered together. **B**. Correlation matrices for the month of April. Note that we do not have mission information in April 2017 and so the data is missing.

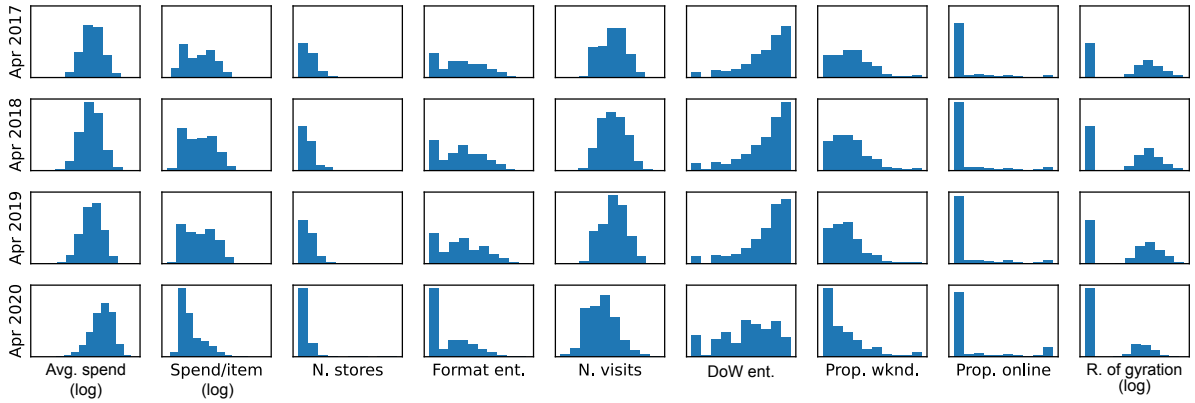


Figure 5.15: Distribution of nine selected metrics (columns) during the month of April over the fours years of records (rows). For anonymisation reasons, the values on the axes are not shown.

In the second group, the number of unique stores visited is highly correlated with the store entropy (yearly: 0.873; Apr.: 0.854) and the format entropy (yearly: 0.733; Apr.: 0.704).

Third, we find that the total number of visits is correlated highly with the mission entropy (yearly: 0.733; Apr.: 0.628; we exclude the year 2017 for which the information is missing) and the day-of-the-week entropy (yearly: 0.710; Apr.: 0.599), at least for the yearly averages though significantly less for the month of April.

The fourth group has the total number of baskets done online and the metric which tells us the proportion of the total visits that online represents (yearly: 0.866; Apr.: 0.848).

Finally, the radius of gyration is positively correlated specially with the total distance travelled (yearly: 0.742; Apr.: 0.712).

5.4.2 Multi-metric analysis

Based on the correlation analysis, we select a subset of nine metrics — the average spend and average spend per item, the visits and the number of stores visited, the format and day-of-the-week entropies, the proportion of online and weekend shops and the radius of gyration — that are either good candidates to represent each of the correlated groups or that are interesting to us for other reasons.

In Figure 5.15, we compare the histogram for these nine measures during the month of April across the four years (we obtain similar results for other months). We can think of each of these histograms like the cross-section at a given point in time for the time-series analysed in Section 5.2.1.

Because the travel and the metrics that measure the basket spend vary over such a large range, we have calculated the logarithm of the average basket spend and the average spend per item. We are also showing the logarithm of the radius of gyration, though we have set the values that were originally zero to one before calculating the logarithm in order to avoid negative infinity. This means that the tall bar on the left side of the radius of gyration histograms corresponds to these households which originally had radii of zero.

We remark that the three top rows look quite similar and that the pandemic makes the

bottom row look the most different. The distribution of the number of spend per item, different stores visits, number of visits, format entropy and radius of gyration are all shifted to the left while the average spend increases and the distribution is shifted to the right. We remember that we took the logarithm for both the average spend and average spend per item so the corresponding shifts are actually quite important.

A very significant change occurs for the proportion of shopping done during the weekend: during the lockdown of April 2020, by far the most common value is zero, indicating that there is now a significant proportion of the population that is not shopping during the weekend.

For the proportion of online shopping, the distribution is bimodal with the most common values being zero or one, though zero is by far the most common. For the majority of the customers sampled, shopping online is either all or nothing. We do see that the number of households that only shop online increased in 2020 compared to the previous years, but we remember that it takes time to see how much online increased due to a lag in the *Retailer*'s capacity to fulfil those orders.

We can see the pairwise comparison of the metrics in the form of scatter plots inside Figure 5.16. Out of the four months which we saw previously, we selected April 2019 to examine what a typical (i.e. non-pandemic) year looks like. Based on the histograms from before (Fig. 5.15) and the scatter plots, we decided not to apply any clustering algorithms because the data does not show any evidence of being composed of distinct clusters, at least not of the kind that can be seen as a separate cloud of points. This however, does not mean that we cannot extract information about different customer habits and we have selected a few panels that caught our attention.

In Panel A, we have the number of visits plotted against the format entropy. The leftmost column corresponds to a zero entropy, and these are the customers that only visit a single format but we can see that this can still lead to a large number of visits. Interestingly, the correlation between the two variables does not seem to be very large, and the customers with the most visits have a mid-range format entropy while the most unpredictable customers that visit all the formats in the same proportion (this would mean a maximal entropy) have a relatively low number of visits.

In Panel B we have the day-of-the week entropy on the X axis while on panel D we encounter again the format entropy; on the Y axis we have respectively the proportion of shopping done during the weekend and the proportion of online shopping.

Both panels look very similar and in particular they have a very interesting structure with curves. The leftmost curve is obtained for the customers that only shop during two days (Panel B) or visit only two different stores (Panel D), then we see the envelope for those that shop 3 days, 4 days, etc. (Panel D: visit 3 stores, 4, etc.)

Next, in Panel C (proportion of online shopping versus number of different stores) we obtain a hyperbolic shape that tells us that the number of stores visited decreases when the proportion of online shopping increases, or vice-versa. This result seems quite intuitive but it is worth noting that the exact relationship that could be inferred as the shape of the curve should be taken with caution, because when a customer shops online they do not actively decide the store

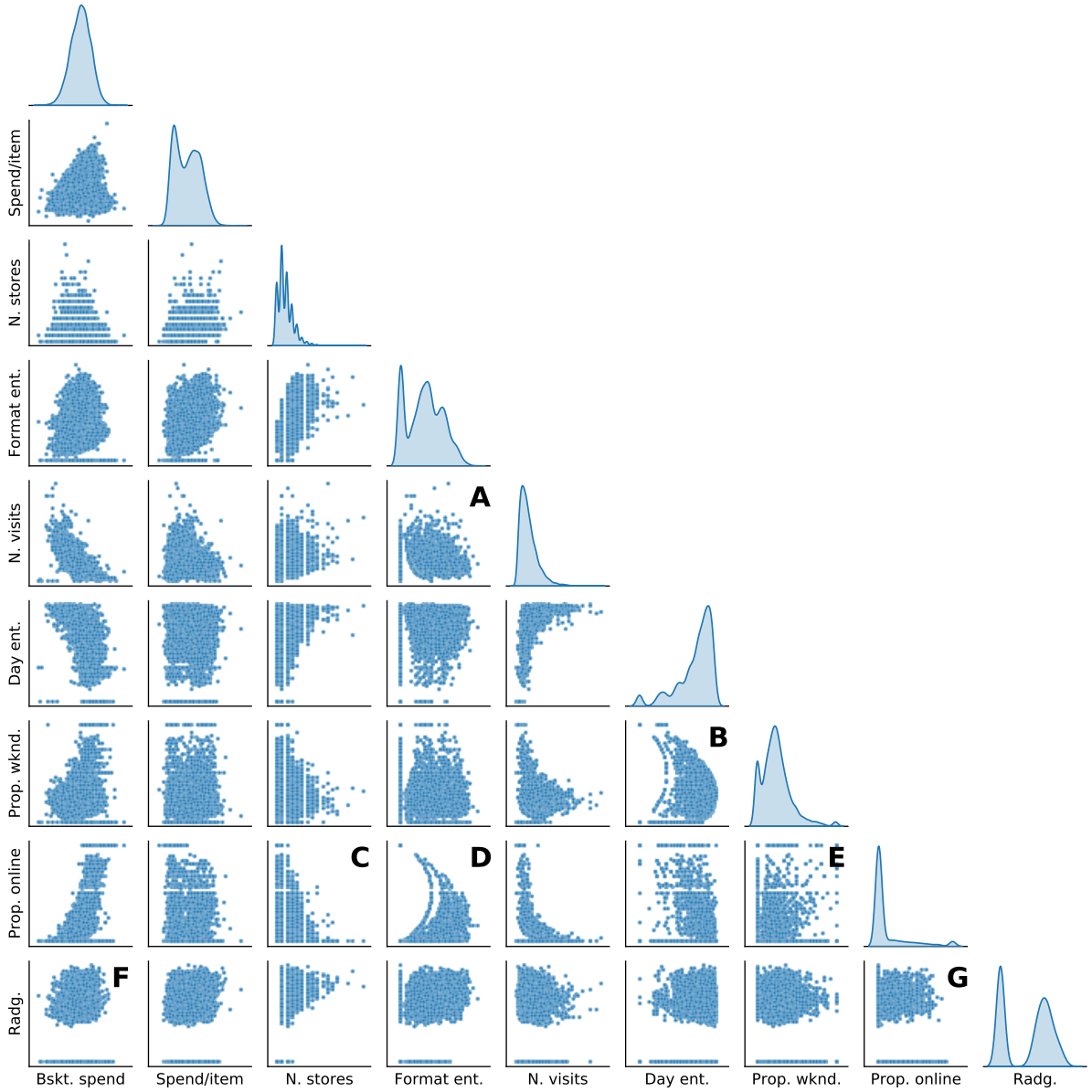


Figure 5.16: Pairwise comparison of the scatter plots belonging to the selected nine metrics during the month of April 2019. The annotated panels are discussed in the main text.

from which the order is served, however this could still manifest as visits to different stores in our data.

In Panel E we have the proportion of weekend shops versus the proportion of online shops. The two diagonals are quite striking: the one following $y = x$ indicates that there is a group of customers for which their weekend and online shopping grow proportionally (we imagine that these could be working households that do some offline shopping during the week and maybe a larger basket online during the weekend); the one following $y = 1 - x$ tells us about a second group for which when either the online or the weekend shopping increase, the other decreases in the same proportion (online shopping during the week and offline baskets during the weekend only).

Panels F and G both have the radius of gyration on the Y axis: the point-clouds which are

pretty flat tell us that this measure of mobility is more or less independent from basket spend and the proportion of proportion online. These two results seem counter-intuitive, but they are supported by the data.

5.4.3 Mobility-based clustering

Having been through the combinations of different metrics, we would like to see if there are different mobility profiles. So far, we have focused on a single point in time (a month or the same month but looked at across the four years) to simplify the comparison of the different metrics. We now take the opposite approach: we will only work with the radius of gyration and what we can change now is the points in time when we look at the values of the metric.

After seeing how the radius of gyration is affected by seasonal patterns (see Figure 5.2), we hypothesise that it would be possible to distinguish these effects and for example classify the customers in terms of an increase (or decrease) in their mobility during the holiday seasons. Furthermore, the way in which people respond to the pandemic in terms of their mobility is probably also something that can be detected.

We use the radius of gyration with the same processing that we described previously (the logarithm and the shift of the zero values). We do not expect this processing to affect how the metric behaves. In particular, we know that this will not affect the large separation between the non-travellers and the rest because the smallest nonzero values are in the hundreds of meters and therefore the logarithm base 10 starts close to 2. As we will see, this separation almost makes the radius of gyration behave like a binary metric that tells us whether the customers shopped in more than one location or not.

It is important to consider what is the reason for this large gap in the radius of mobility, and we are confident that it is an effect of being hindered by a resolution limit both in the temporal and spatial dimensions.

The temporal limit is easy to grasp: the majority of customers are not shopping more than once or twice every week, and only those shops that are made offline count towards the mobility. This is unlike the situation for GPS tracking where it is normal to have tens or even hundreds of data points on a daily basis.

The spatial limit comes from the fact that we sample the possible locations in all of the UK with the set of stores, roughly corresponding to 3,000 locations. Having a countable and relatively small number of unique places where a customer can be is very useful, for example, for the flow reconstruction problem, but it hurts us when we calculate the radius of gyration. To make matters worse, we have the problem of the overlapping coordinates and, if a customer only shops in a large store and the associated petrol station, their radius of gyration will be zero.

Leaving these problems aside, let us now investigate if there are different types of mobility profiles. We still select a monthly frequency to keep the number of features manageable and more importantly to mitigate the amount of customers that have a zero radius of gyration (which obviously happens more often the shorter the time window is).

We embed the variables into two dimensions because reducing the dimensionality of the problem usually makes the computational tasks easier, but also because we will experimenting

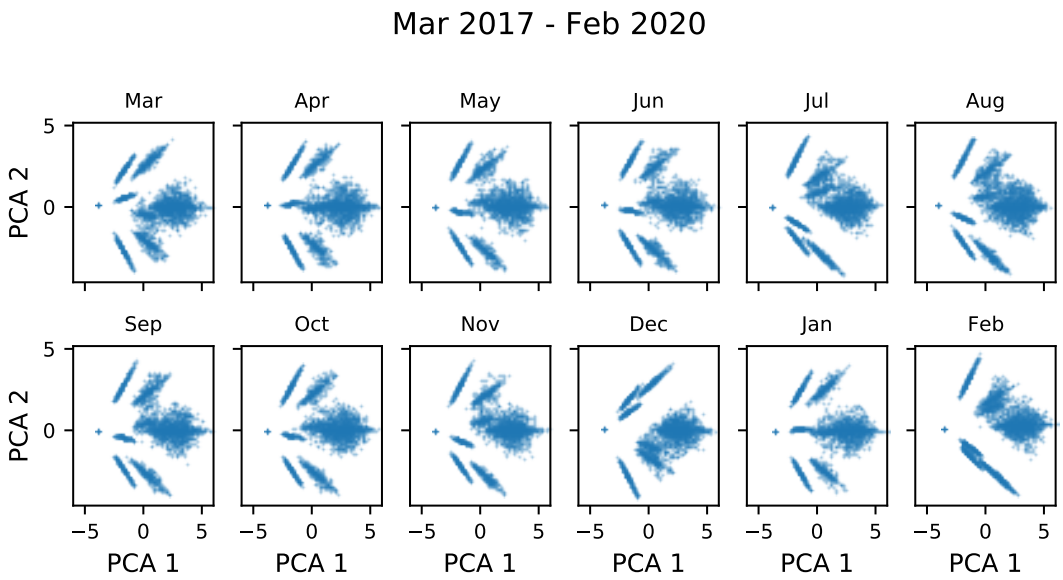


Figure 5.17: PCA embedding of the set of three variables that correspond to any given month during the span of 2017 to 2019.

with different groups of months which means a variable number of features and the embedding allows us to do a consistent treatment always in terms of two or three coordinates. We present the results obtained using a linear embedding with principal components analysis (PCA). We also experimented with a nonlinear embedding but our algorithm of choice, Isomap, does not appear to be well suited to the almost binary behaviour of the radius of gyration and we do not include the results.

Linear embedding (PCA)

PCA is one of the most popular techniques to transform a high-dimensional set of features or variables into a reduced number of components that are orthogonal to each other and also preserve the largest amount of variance. If we arrange the input data (n p-dimensional data points) into the matrix $\mathbf{X}$, we obtain the projection

$$\mathbf{T} = \mathbf{X}\mathbf{W}\,, \tag{5.32}$$

where W is a $p \times q$, $q \leq p$ matrix of weights. Below, we select $q \in \{2, 3\}$.

We first look at the years before the pandemic (covering March 2017 to Feb 2020) and we start by viewing the months individually. For the three years, we have three sets of variables for any given month (that is, if we select April, we employ the months of April 2017, April 2018 and April 2019). The data matrix will therefore hold $p = 3$ columns and we set $q = 2$ to be able to visualise the results. The projected points in $\mathbf{T}$ can be seen in Figure 5.17.

The separate clusters are essentially the possible combinations of the variables being zero or nonzero (one of them, two of them or all three of them). The first principal component — which explains the majority of the variance — lines up with the lonely point on the left hand

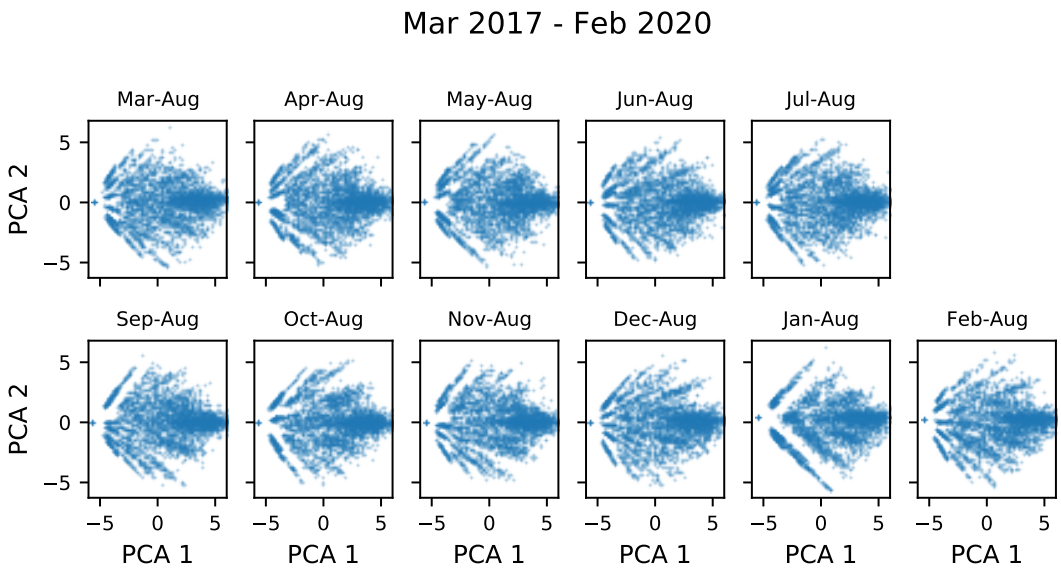


Figure 5.18: PCA embedding of the set of six variables that correspond to pairs of months (one of which is selected to be August) during the span of 2017 to 2019.

side which corresponds comes from the household that have a zero radius of gyration during each of the 3 months, and those that have nonzero and high radii the 3 months. We cannot use this in any particularly useful way: we are essentially seeing what we said before which is the binary nature of the radius of gyration which almost behaves as being on or off.

Next we plot pairs of months. In Figure 5.18, we show the $\mathbf{T}$ projection ($q = 2$) when the original data corresponds to the month of August for the three years as well as a second month, giving a total of $p = 6$ variables (three for August and three for the second month). We can see how we are getting a similar behaviour as before, with the clusters on the left hand side corresponding to the projection of the points for which one or more of the months are zero. We selected August because we associate it with the travelling during the summer holidays, but the results are basically indistinguishable from the other months.

Until now, we have avoided including the fiscal year 2020 (spanning from March 2020 to February 2021) because it does not look like the others (as attested in the histogram comparison in Fig. 5.15). To analyse this year, we find it more informative to compare the projection of the twelve months that make up each year. From these sets of $p = 12$ variables, we compute the principal components, this time selecting $q = 3$ so that the projection is three-dimensional.

As before, the first principal component accounting for the largest variance lines up with the zero radii points on one side, and those with nonzero and large radii on the other (the left-right axis that we have already encountered in Figures 5.17 and 5.18) and does not interest us particularly. Instead, we analyse the second and third components (second and third columns in $\mathbf{T}$) which can be found in Figure 5.19 (this view is orthogonal to the plots we have seen above). Since we already know that the points that are given a negative first component are those for which some of the months have a zero radius of gyration, we have discarded them from this new figure. Quite remarkably, we can now distinguish fairly easily separate clusters, and this is true for the four years.

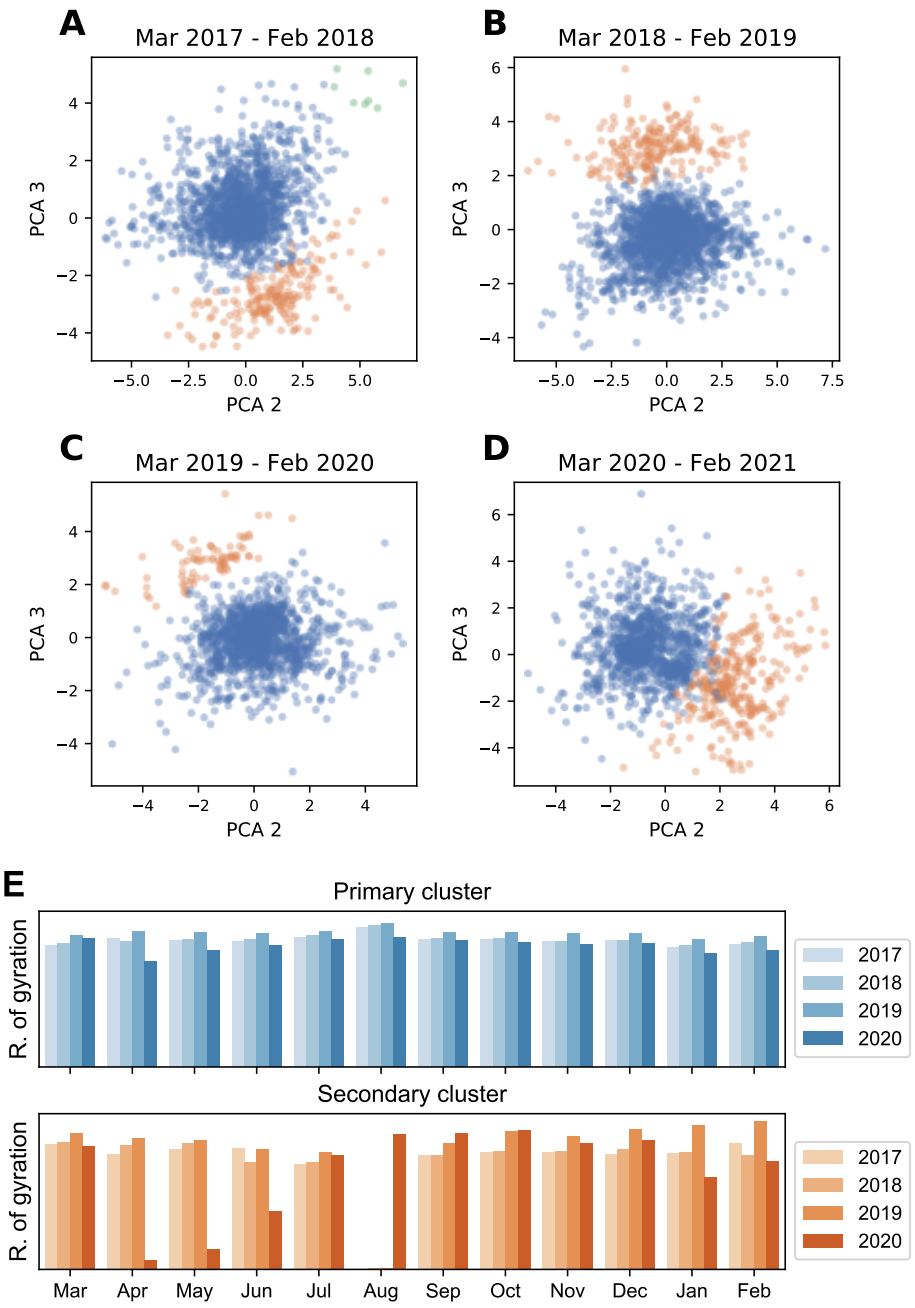


Figure 5.19: **A** to **D**: Second and thirds principal components calculated from the set of twelve variables that make up each year (respectively 2017 to 2020). We have discarded the points with a negative first component, and the colours correspond to clusters detected using spectral clustering. **E**: having classified the points into two clusters, we now give the average radius of gyration of each group during the different months.

We run spectral clustering with 2 components (3 for the case of 2017 though a cluster is only composed of 8 points) to the points that are represented in each of the panels in Figure 5.19, and we detect the central or primary cluster which has the households with the most mobility in the 12 months (in blued), and a secondary cluster (in orange).

To probe the composition of these two groups, we calculate the average radius of gyration within the group for each month (Figure 5.19E). For the years 2017 to 2019 included, this second cluster is quite distinctly composed of those users that have a zero radius during the month of August (notwithstanding a small amount of contamination). Interestingly, the separate cluster for 2020 is not the month of August but instead a majority of zero radii during the months of April and May (the average becomes lower when there are many zero values; due to the large gap we know that the values do not go down continuously).

This results shows that indeed there is a signal that can be detected, and year on year the month of August is a special month for many households. This is almost surely related to the summer holidays, though the result is surprising because it is the opposite of what we were expecting. We thought that the month of August would lead to increased travelling and therefore an increased radius of gyration. What we are finding is the opposite behaviour and it would be tempting to think that these customers are travelling abroad, but this would not be the whole story because they appear in our data and therefore they are shopping during this month. Another possible explanation is that these customers do go on vacation somewhere different and once at their destination they are doing all their shopping at a single shop, which would manifest as a zero radius of gyration.

For the year 2020, this has a simpler explanation: the drop for this group of customers coincided with the first lockdown and these are the customers that are not shopping in more than one location.

Chapter 6

Assessing navigation ability for early detection of Alzheimer's Disease

For the final chapter, we change the application and we study a dataset concerned with navigation in the context of dementia and Alzheimer's Disease. Though we do not encounter spatial networks directly, the origin-destination matrices that we have seen before in the context of spatial networks will be immensely useful to define new measures of navigation ability.

6.1 Introduction

Dementia affects millions of people around the world. The Alzheimer's Society — a UK charity which funds research and does advocacy — estimates that in 2020 there were 36 million people with dementia worldwide (some 850,000 in the UK) and by 2050 some 115 million will be affected (almost 1.6 million in the UK) [115].

Alzheimer's disease (AD) is the most common cause of dementia, accounting for some 60 to 70% of cases [80]. It initially manifests itself as forgetfulness of recent events and disorientation, progresses into behavioural changes like trouble remembering people's names or problems with language, and in the final stages the memory problems worsen while bodily functions (e.g. walking, continence, swallowing) are gradually lost, eventually leading to death. According to the World Health Organization, dementia is currently the seventh leading cause of death among all diseases [80].

There is no known treatment that can cure AD, though early detection can still be useful in order to provide forms of palliative care that can alleviate some of the symptoms. Of the cognitive fingerprints of AD, episodic memory deficits are currently used as the gold standard for diagnosis [17]. However, memory declines with normal ageing and it can be difficult to distinguish age-related changes from the memory deficits caused by pathophysiological processes [17].

In this setting, spatial disorientation could be an overlooked cognitive marker which provides key advantages because it can be seen in early-onset AD with the first pathophysiological changes can taking place up to 20 years before diagnosis [20]. There is also mounting evidence that spatial navigation could be impacted before episodic memory deficits emerge [18].

An effective use of spatial navigation has its own challenges. Indeed, navigation varies naturally among individuals: age is a strong predictor of declining ability and sex plays a significant role too (with males outperforming females). Recently, navigation ability has been studied on a global scale across 57 countries [20] and found to be correlated with the development of a country (as measured, for example, using GDP per capita). Furthermore, navigation ability is influenced by environmental factors such as the layout of the city where an individual resides [19]. To leverage spatial navigation for early AD detection, one must account for the inter-individual variability to be able to determine what can be considered abnormal.

Researchers at University College London and the University of East Anglia have developed a mobile game called Sea Hero Quest (SHQ) which has been downloaded millions of time around the world. Players can voluntarily answer demographic questions about themselves, and the data thus collected has the potential to serve as a massive collection which includes the common and uncommon trajectories that people take to get somewhere (albeit in a virtual environment). "Data from the game could be used to create the first population benchmarks for healthy navigation abilities across ages, gender and countries. In turn, these benchmark scores could enable the development of easy-to-administer, sensitive spatial navigation tools that are validated against population-level benchmarked scores and related to real-life navigation problems that patients might encounter." [17]

It is to the objective of developing personalised benchmarks that we set out to contribute. With a colleague, we introduce novel metrics (she contributed the smoothing of the trajectories and the average curvature and boundary-affinity metrics while the others were my idea) that can be applied to SHQ gameplay data — and we hope in other settings too. I devised how to use the new measures to study the ability of two clinical cohorts and I also propose the methodology for comparing individual performances against a normative population.

6.1.1 Sea Hero Quest

Created in 2016, the Sea Hero Quest mobile game has been played by 4.3 million people all over the world giving life to a dataset of unprecedented scale in dementia research [120].

In the game, one plays the role of a sea captain completing adventures in order to recover the lost memories of his father who suffers from dementia. Players steer a boat based on the visual cues that they can see around them (see Figure 6.1A).

The game consists of two types of levels wich are designed to test two different mechanisms used in navigation: wayfinding levels and path integration levels. In the wayfinding levels, a player is shown a map with numbered checkpoints that have to be visited in order; the map then disappears and the player must steer the boat and find their way. The geometry of the levels varies greatly, with lower levels being quite open and having checkpoints that are easy to locate from afar, and higher levels being narrower and with checkpoints hidden by the terrain.

In path integration levels, the player has to navigate to the end of a river to find a flare gun. When they have reached the gun, the boat turns around and the player is asked to select from three possible directions the one which points to the starting point.

For the present analysis, we focus on wayfinding levels. We study the trajectories generated

Allele	General pop. Frequency (%)	Pop. with AD
Apo-E2	8.4	3.9
Apo-E3	77.9	59.4
Apo-E4	13.7	36.7

Table 6.1: Frequency of the different alleles in the general population and in the sub-group that has developed AD. Adapted from Liu et al. [60]

by the players in terms of their length, shape, etc., but also in terms of their relationship to the geometry of the level (characterised by the position of the checkpoints and the boundaries). Of the wayfinding level, the most important and the ones which we study are Levels 6, 8 and 11 (left to right in Figure 6.1, panels B-D). These three levels are encountered early enough in the progression of the game that a large number of at-home players have attempted them and we have a sizeable population from which we build our normative benchmarks. By contrast, higher levels have been played in smaller numbers and they have also not necessarily been attempted by the clinical patients.

Importantly, we only analyse the first attempt made by an individual to complete a particular level. This we do in order to avoid the influence of learning, which albeit interesting, will introduce yet another dimension along which spatial navigation ability can differ amongst individuals.

6.1.2 At-genetic-risk and dementia clinical groups

Though the causes of Alzheimer's disease are yet to be fully characterised, it is understood that the main genetic risk factor of late-onset AD is the Apolipoprotein E (APOE) $\epsilon 4$ allele. Apolipoprotein E is involved in the transport of lipids and injury repair in the brain [60]. The gene regulating it is has three main variants: the $\epsilon 4$ allele increases the risk of AD compared to the neutral allele $\epsilon 3$, whereas $\epsilon 2$ carriers are at a lower risk. We can find the population-wide and AD frequencies of the variants in Table 6.1.

Taking $\epsilon 3$ homozygotes as a baseline, caucasic individuals with one copy of $\epsilon 4$ have a 2.6 or 3.2 higher risk (respectively for the second allele $\epsilon 2$ and $\epsilon 3$) while having two copies increases the risk by a factor of 14.9 [60].

In a study of at-genetic-risk patients, Coughlan et al. [18] compare the Sea Hero Quest gameplay data of a population of $\epsilon 3\epsilon 3$ carriers and $\epsilon 3\epsilon 4$ carriers. They establish that the $\epsilon 4$ allele has an effect on the distance travelled but also that distance travelled has a better predictive power to infer the APOE genotype than a memory task. This study proves the potential to employ SHQ in a clinical setting but there is still a great deal of information to extract from the trajectories, and a need remains to "establish personalized normative measures to accurately assess spatial disturbances".

Our work is meant to be in part a follow-up to the Coughlan study. Our collaborator at the University of East Anglia shared the spatial navigation data of the genotyped cohort from the study. Participants were all between 50 and 75 years old when they were recruited

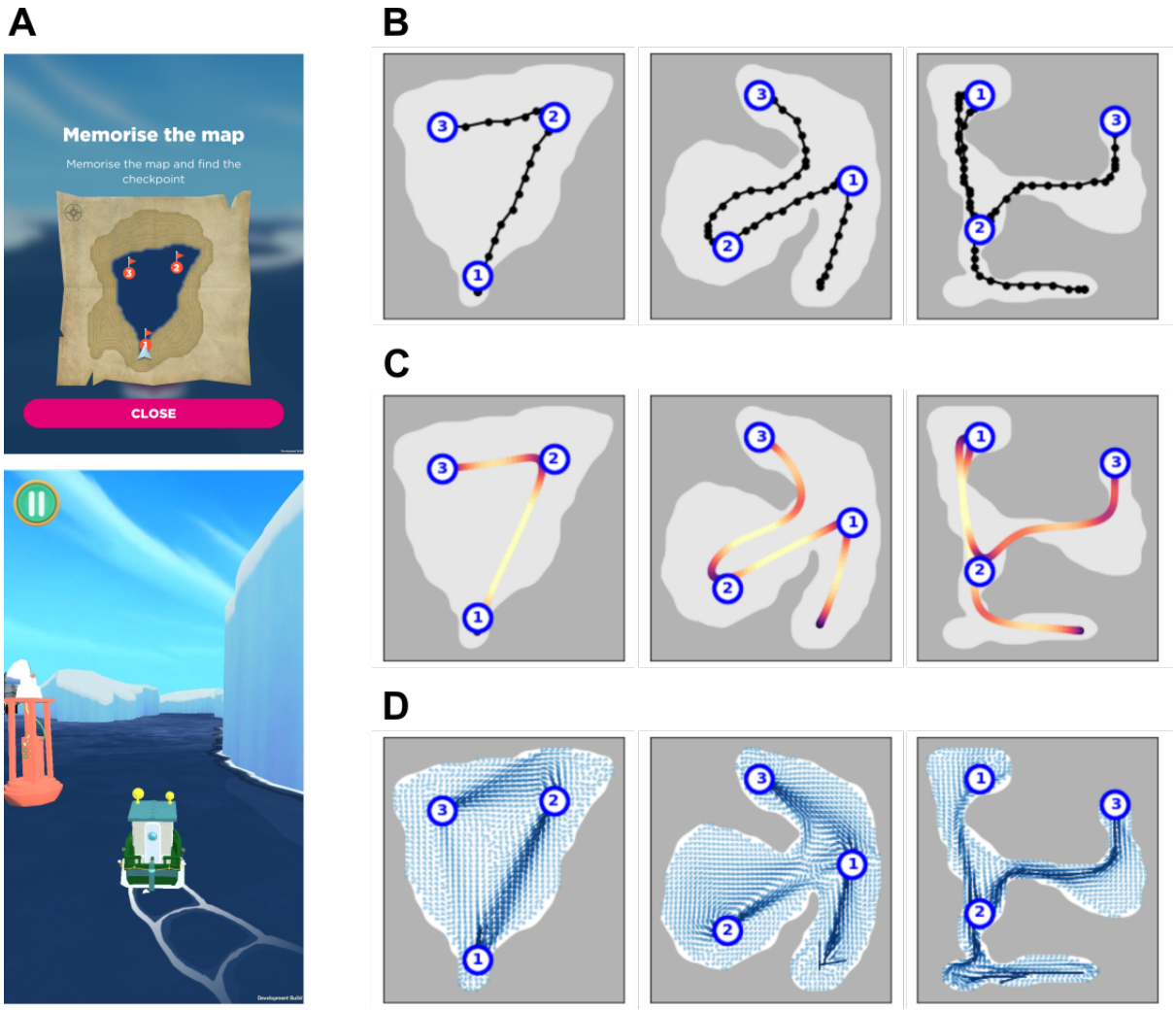


Figure 6.1: **A**. Screen captures taken from the Sea Hero Quest game (provided by Michael Hornberger and Hugo Spiers). **B.** Wayfinding trajectories as originally recorded. From left to right, levels 6, 8 and 11. The trajectories start from the bottom right and the checkpoints should be visited in the order indicated. **C.** Reconstructed smooth interpolation of the trajectories, coloured using the velocity. **D**. Mobility fields computed from the normative population (in this case males aged 50 to 59). Panels B-D were drawn by Uzu Lim with the data and the OD matrices that we generated for her.

Level	Females		Males	
	N	%	N	%
6	31843	52.2	29133	47.8
8	27962	51.0	26870	49.0
11	22623	50.1	22490	49.9

Table 6.2: Composition of the normative dataset for the three wayfinding levels studied.

between February and June 2017. After passing a screening for psychiatric or neurological disease, substance dependence and normal or corrected-to-normal vision, their APOE genotype was determined from a non-invasive saliva sample.

The resulting cohort is composed of three groups, the group of $\epsilon 3\epsilon 3$ carriers ($n = 28$),[1] another of $\epsilon 3\epsilon 4$ carriers ($n = 31$), and a third group of homozygous $\epsilon 4\epsilon 4$ carriers ($n = 5$). This last group is so small that it has hard to draw statistically valid conclusions and we exclude it from some parts of our analysis. Carriers of the third allele $\epsilon 2$ were also excluded from the original study and we do not have data related to them.

Further to this cohort, we also have access to the SHQ data generated by a group of $n = 21$ patients from a different study, aged 53 to 79 years, and that have been diagnosed with Alzheimer's disease or with amnestic Mild Cognitive Impairment (aMCI) — a disorder which manifests primarily through memory loss and can evolve into AD. The exact progression from MCI to AD is not fully determined, but patients with MCI show spatial navigation impairments much like patients with early AD [17]. For the purpose of our analysis, we treat both types of patients as a single group.

6.1.3 The normative dataset

To establish our benchmark population, henceforth called normative, we focus on the at-home players that voluntarily provided their gender, age and country of residence. Given that the study of the clinical patients is our primary focus, we select for our normative population the subset of users residing in Great Britain and 50 to 80 years in age.

Naturally, this self-reporting can lead to false information. Evidence of this can be seen in the abnormally high number of users who report that they are 99 years old (the maximum age that can be entered). However, limiting ourselves to users between 50 and 80 years old is probably far enough from the extremes which are more likely to be problematic and we do not expect the self-reporting to lead to a noticeable systematic error.

There are between 45 and 60 thousand recorded attempts in Great Britain for each of the levels studied. The exact numbers can be found in Table 6.2.

[1] The study of Coughlan et al. reports one more participant in this group, but in private communications we were told that this person had failed to return a questionnaire and had their demographic information missing. We therefore excluded this patient from the present analysis.

6.2 Measures of navigation ability

We have mentioned above that we study three SHQ wayfinding levels. The levels are different in their geometries but they can all be analysed with the same tools. Therefore, without loss of generality, we introduce the navigation ability measures for a single level.

The position of the boat in Sea Hero quest is sampled at a constant rate of 2 Hz. We denote the recorded trajectory which is τ time-steps long as $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \cdots \boldsymbol{\gamma}_\tau)$. In the mobile game, the user can steer the boat anywhere they like on the water but the recorded trajectory — probably because of data storage concerns — encodes the positions on a discrete two-dimensional grid which gives us to the coordinate pairs $\boldsymbol{\gamma}_n \in \mathbb{N}^2 \;\; \forall n \in \{1, \cdots, \tau\}$.

The grid is quite coarse and the encoding introduces a rounding error which is not insignificant; indeed, the trajectories look "boxy" and they will be too coarse for some purposes (see Figure 6.1B). We calculate a smooth interpolation $\boldsymbol{\sigma}$ which is a reconstruction for the trajectory which could have been recorded with more detail. To define it, we first compute a piece-wise linear interpolation of the original trajectory as

$$\mathbf{x}_n^{(k)} = \boldsymbol{\gamma}_n + \frac{k}{r}\left(\boldsymbol{\gamma}_{n+1} - \boldsymbol{\gamma}_n\right) \,, \quad \text{for } n < \tau, \;\; k \in \{0, \cdots, r-1\} \,, \tag{6.1}$$

$$\mathbf{x}_\tau^{(0)} = \boldsymbol{\gamma}_\tau \,. \tag{6.2}$$

The parameter $r \in \mathbb{N}$ controls the number of points that we use to reconstruct the trajectory, and it would seem that $r = 3$ generates plausible-looking and smooth trajectories without incurring too great a cost in terms of the increased length of the trajectory.

Next, we order the $T = r(\tau - 1) + 1$ auxiliary points so that they parametrise the trajectory in the right order $\left(\mathbf{x}_1^{(0)}, \cdots, \mathbf{x}_1^{(k-1)}, \mathbf{x}_2^{(1)}, \cdots, \mathbf{x}_2^{(k-1)}, \cdots, \mathbf{x}_\tau^{(0)}, \cdots, \mathbf{x}_\tau^{(k-1)}, \mathbf{x}_\tau^{(0)}\right)$, and we apply a one-dimensional Gaussian filter with standard deviation ν (a convolution with a Gaussian kernel) separately to each coordinate. We use reflected boundary conditions. We denote the smooth trajectory by $\boldsymbol{\sigma} = (\boldsymbol{\sigma}_1, \cdots, \boldsymbol{\sigma}_T)$. We can see the interpolated trajectories obtained with $r = 3$ and $\nu = 5r/3$ as the black solid curves in Figure 6.1B. In the future we hope to carry out a robustness analysis to explore the effect of selecting other parameters.

Having introduced the notation for the trajectories, we define a number of measures which we hope can characterise the spatial navigation ability of an individual. We distinguish between absolute measures and relative measures. The former can be computed given a single trajectory while the latter make reference to a set of trajectories (which would normally be taken from the normative dataset).

6.2.1 Absolute measures

To measure the distance travelled, otherwise known as the length of the trajectory, we use the smooth interpolation,

$$L[\boldsymbol{\sigma}] = \int_0^T ||\boldsymbol{\sigma}'(t)||_2 \, \mathrm{d}\, t \,. \tag{6.3}$$

The duration of an attempt is recorded outside the trajectory as additional metadata. Together with the trajectory length, it was employed by Coughlan et al. [18]. Note that a player in SHQ controls the speed of the boat and as such the same navigation path can be completed in different times. The duration is correlated but nonetheless different to a simple multiple of the trajectory length.

New measures

We introduce the average curvature of the trajectory, calculated using

$$C[\boldsymbol{\sigma}] = \frac{1}{\tau}\int_0^T \left\| \frac{\mathrm{d}\,\hat{\boldsymbol{\sigma}}'(t)}{\mathrm{d}\,t} \right\| \mathrm{d}\,t, \text{ where } \hat{\boldsymbol{\sigma}}'(t) = \frac{\boldsymbol{\sigma}'(t)}{\|\boldsymbol{\sigma}'(t)\|}\,. \tag{6.4}$$

The name curvature might imply that this is a signed measure but this is not the case for the definition given above. We care about the absolute sum of the tangent angles to measure the overall "curviness" of a trajectory.

Coughlan et al. [18] report that $\epsilon 4$ carriers show a lower average distance to the border, and this can indeed be observed for some of the trajectories, though the study does not provide quantitative evidence for this finding. To address this, we define a measure called boundary affinity. If we let $M \subset \mathbb{R}^2$ denote the land-mass which the player cannot navigate into,[2] and $d(\boldsymbol{\sigma}(t), M)$ is the distance between the trajectory at time t and the nearest point on the boundary, we define

$$B[\boldsymbol{\sigma}] = \frac{1}{T}\int_0^T \frac{1}{1+\exp(-f(d_t))}\,\mathrm{d}\,t, \text{ where } d_t = d(\boldsymbol{\sigma}(t), M)\,, \tag{6.5}$$

where f is a rescaling function which we are free to choose. We select a function parametrised with two radii, $f(d) = \alpha\frac{2d-(r_{\text{out}}-r_{\text{in}})}{r_{\text{out}}-r_{\text{in}}}$. By experimentation, we find that $r_{\text{in}} = 1.5$ and $r_{\text{out}} = 4$ are appropriate for levels 6 and 8, while for the narrower level 11 we choose $r_{\text{in}} = 1$ and $r_{\text{out}} = 2$. For all the levels $\alpha = 4$.

Visiting order

Given the set of checkpoints $C = \{C_1, \cdots, C_e\} \subset \mathbb{N}^2$ which are part of the geometry of a level, we compute the order in which a trajectory $\boldsymbol{\gamma}$ visits them (the discrete non-smooth trajectory works well in this case). We do this in two steps. First, we compute for each location whether a checkpoint has been visited and is less than R pixels away[3] and we store its index,

$$\boldsymbol{\gamma}_n \mapsto \begin{cases} s & \text{if } \mathbb{1}_{\mathrm{B}_R(C_s)}(\boldsymbol{\gamma}_n) = 1\,, \\ 0 & \text{otherwise.} \end{cases} \tag{6.6}$$

[2] Like the trajectory $\boldsymbol{\gamma}$, the land-mass is originally encoded on top of the discrete grid leading to an unnecessarily coarse definition of the boundary. To recover a more faithful representation, we manually identify the contour of points which define the inner boundary of the level and we compute a smooth interpolation of this set of points.

[3] For a checkpoint to be considered visited, players must cross get inside a circular boundary which is visible in the game. Unfortunately, our colleagues did not know what the radius used by the game developers was but we carried out experiments and found $R = 3$ to give sensible results.

Next, we group the repeated nonzero values and select only one element for each set of contiguously repeated indices, discarding the zeros. For example, the sequence $(0, 1, 1, 0, 2)$ becomes $(1, 2)$.

After computing the order in which the checkpoints were visited, one can easily determine whether it corresponds to the order expected or not. We express this using a binary flag which we call the *visiting order correctness*. This measure, simple as it is, has not been used in previous studies and it plays a very important role as we will see.

6.2.2 Relative measures

Relative measures characterise a trajectory by means of a comparison with other attempts, and they are all newly proposed here. Inspired by our research in mobility networks, we had the idea of using origin-destination matrices. In this, we actually take advantage of the coarse grid over which $\boldsymbol{\gamma}$ is recorded, because the pixels that make up the grid are easily numerable and can be thought as locations in the OD framework.

For a level that has width w and height h, let $G = \{(x, y) \in \mathbb{N}^2 : 0 \leq x \leq w, 0 \leq y \leq h\}$ be the set of pixels (discrete valid positions). We order the pairs in G lexicographically: $G = \{\mathbf{G}_1, \cdots, \mathbf{G}_\ell\}$ where $\ell = w \cdot h$ and $(x, y) = \mathbf{G}_{wy+x}$.

Consider N (discrete) gameplay trajectories, $\Gamma = \{\boldsymbol{\gamma}^{(1)}, \cdots, \boldsymbol{\gamma}^{(N)}\}$, each with gameplay time $\tau^{(1)}, \cdots, \tau^{(N)}$. The *origin-destination matrix*, or *OD matrix*, is the $\ell \times \ell$ matrix whose (i, j)-th entry holds the number of times the individuals in Γ went from G_i to G_j in a single timestep,

$$\mathrm{OD}[\Gamma]_{i,j} = \left| \{(\boldsymbol{\gamma}, t) : \boldsymbol{\gamma}_t = i, \boldsymbol{\gamma}_{t-1} = j\} \right| . \tag{6.7}$$

Given a second set of gameplay trajectories Γ' with N' elements, we compare Γ and Γ' by taking the difference of the two (normalised) OD matrices

$$\Delta[\Gamma, \Gamma'] = \left\| \frac{\mathrm{OD}[\Gamma]}{N} - \frac{\mathrm{OD}[\Gamma']}{N'} \right\| . \tag{6.8}$$

In the definition above, we take either the Frobenius norm or the supremum norm (L^∞ norm) for the matrix norm, which we denote by Δ_F and Δ_∞ respectively.

The use case that we are most interested in is when $\Gamma' = \{\boldsymbol{\gamma}\}$ is composed of a single trajectory, and Γ is the set of attempts made by a representative population with matching age and gender. Δ_F and Δ_∞ are the first two of our relative measures.

The *matching-sum* measure Σ is defined by summing all the entries on $\mathrm{OD}[\Gamma]$ that correspond to nonzero entries of $\mathrm{OD}[\{\boldsymbol{\gamma}\}]$,

$$\Sigma[\boldsymbol{\gamma}, \Gamma] = -\frac{1}{N \cdot \tau} \sum_{\mathrm{OD}[\{\boldsymbol{\gamma}\}]_{i,j} > 0} \mathrm{OD}[\Gamma]_{i,j} . \tag{6.9}$$

Intuitively, the sum will be large when $\boldsymbol{\gamma}$ follows a "popular" trajectory that is commonly taken amongst the normative population. The minus sign is included so that a trajectory that goes through common locations scores lower than a trajectory that does not; this makes the matching-

sum measure consistent with how the others behave. Looking back, we could instead take the inverse which feels like a more natural transformation; a positive score that is either small or large feels easier to interpret. However, we need not worry of this being an issue for our analysis and the minus sign will lead us to the same conclusions because the relevant comparisons that we do have taken into account the percentile ranking of individual scores (see Section 6.4.2).

We are also interested in the average direction of travel at each location. Similarly to Mazzoli et al. [66], we define the *mobility field* $\mathbf{F}$ over the grid G by summing all the transitions in Γ happening at a location across a single timestep

$$\mathbf{F_i} = \frac{1}{N} \sum_j \frac{\mathbf{G}_j - \mathbf{G}_i}{||\mathbf{G}_j - \mathbf{G}_i||} \, \mathrm{OD}[\Gamma]_{i,j} \, . \tag{6.10}$$

We use the mobility field in the definition of our last relative measure, dubbed the *mobility functional*, and which is the discrete line integral

$$\Phi[\boldsymbol{\gamma}, \Gamma] = -\frac{1}{\tau} \sum_{t=1}^{\tau-1} \langle \mathbf{F}_t, \boldsymbol{\gamma}_{t+1} - \boldsymbol{\gamma}_t \rangle \, , \tag{6.11}$$

where $\mathbf{F}_t$ is the mobility field at $\boldsymbol{\gamma}_t$, that is, $\mathbf{F}_t = \mathbf{F}_i$ where $\mathbf{G}_i = \boldsymbol{\gamma}_t$. As before, the minus sign makes the mobility functional large when the trajectory $\boldsymbol{\gamma}$ is aligned with common transitions.

6.2.3 Moving age window for the normative population

The relative features require us to consider what is the best way to select the normative set Γ against which a given trajectory $\boldsymbol{\gamma}$ is compared. Let us suppose that $\boldsymbol{\gamma}$ is the attempt made by an individual whose age is a and whose gender is $g \in \{\mathrm{Male}, \mathrm{Female}\}$.

Based on a suggestion by our medical colleagues, we initially divided the normative SHQ players into age cohorts 50-59, 60-69, etc., and matched the individual's gender and age with the group containing a. However, we felt that this was too arbitrary and we decided that a moving age window was more appropriate: we select the set of players aged $[a-5; a+5]$ (note that this covers 11 years but it centres the age group). Alternatively, it might seem justified to select only those players whose age is exactly a but this has the disadvantage that we get a smaller set Γ and it leads to noisier results.

Let us denote the normative trajectories in the moving window by $\{\Gamma^{(a-5)}, \cdots, \Gamma^{(a+5)}\}$. While the moving window is useful because we get a sizeable population which helps with the noise, it includes a varied population that might have quite different navigation abilities due to their age difference. We can limit this by giving a weight to the contribution for each year, weighting the central years around a more heavily. This is the approach we follow, and we select the weights according to a Gaussian centred around a and standard deviation s. Formally,

$$\mathrm{OD}[\Gamma] = \sum_{d=a-5}^{a+5} w_d \frac{\mathrm{OD}[\Gamma^{(d)}]}{|\Gamma^{(d)}|} \, , \tag{6.12}$$

Level	Gender	Completed correctly	Percentage (%)
6	F	24230	76.1
	M	24875	85.4
8	F	9328	33.4
	M	11880	44.2
11	F	16402	72.5
	M	18800	83.6

Table 6.3: Number of trajectories in the normative dataset that were completed visiting the checkpoints in the correct order. Note that Level 8 has lower rates of success because there is a left turn that is hidden by the terrain.

where

$$w_d = Z \frac{1}{s\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{d-a}{s}\right)^2\right], \tag{6.13}$$

and Z is chosen so that $\sum_{d=a-5}^{a+5} d_y = 1$. We empirically find good results with $s = 2$ though we also observed that the results are not too sensitive to the value selected as long as the distribution of weights is not too flat.

6.3 Visiting order separation

We begin by applying our metrics to the normative dataset as a first step to establish their utility in distinguishing the healthy trajectories from those that are obviously problematic. In this setting, we find that the visiting order correctness (VOC) is a simple but also very useful measure because it equips us with a zero-th order criterion which is easy to work with: a trajectory is more likely to be indicative of troubled spatial navigation if the visiting order for the level considered is incorrect. (The converse is not necessarily true as we will see shortly.) Indeed, failing to visit the checkpoints in the correct order is an indication of difficulty integrating the initial map, having trouble steering the boat in direction wanted, or misremembering the order in which the checkpoints must be visited.

Using the VOC, we can separate the trajectories into two groups and describe any differences found across them. For the set of normative players in Great Britain and aged 50 to 80 years, we can find the percentage of attempts completed with a correct order in Table 6.3. Roughly 80% of the players complete Levels 6 and 11 in correctly, but less than half of the males and one in three of the females complete Level 8 correctly. This is an interesting finding because Level 8 does not look particularly difficult from the point of view of the map — the level is quite open which tends to make navigation easier — but the problem seems to lie in a left turn that needs to be made after reaching the first checkpoint, and which the narrow screen makes it difficult to see so that players must remember the map correctly to find their way.

Given that the VOC captures at a very simple level a problematic aspect of navigation, we would hope that the other measures penalise the trajectories that have an incorrect visiting order. Panel A in Figure 6.2 shows the median score and the inter-quartile spread of the different

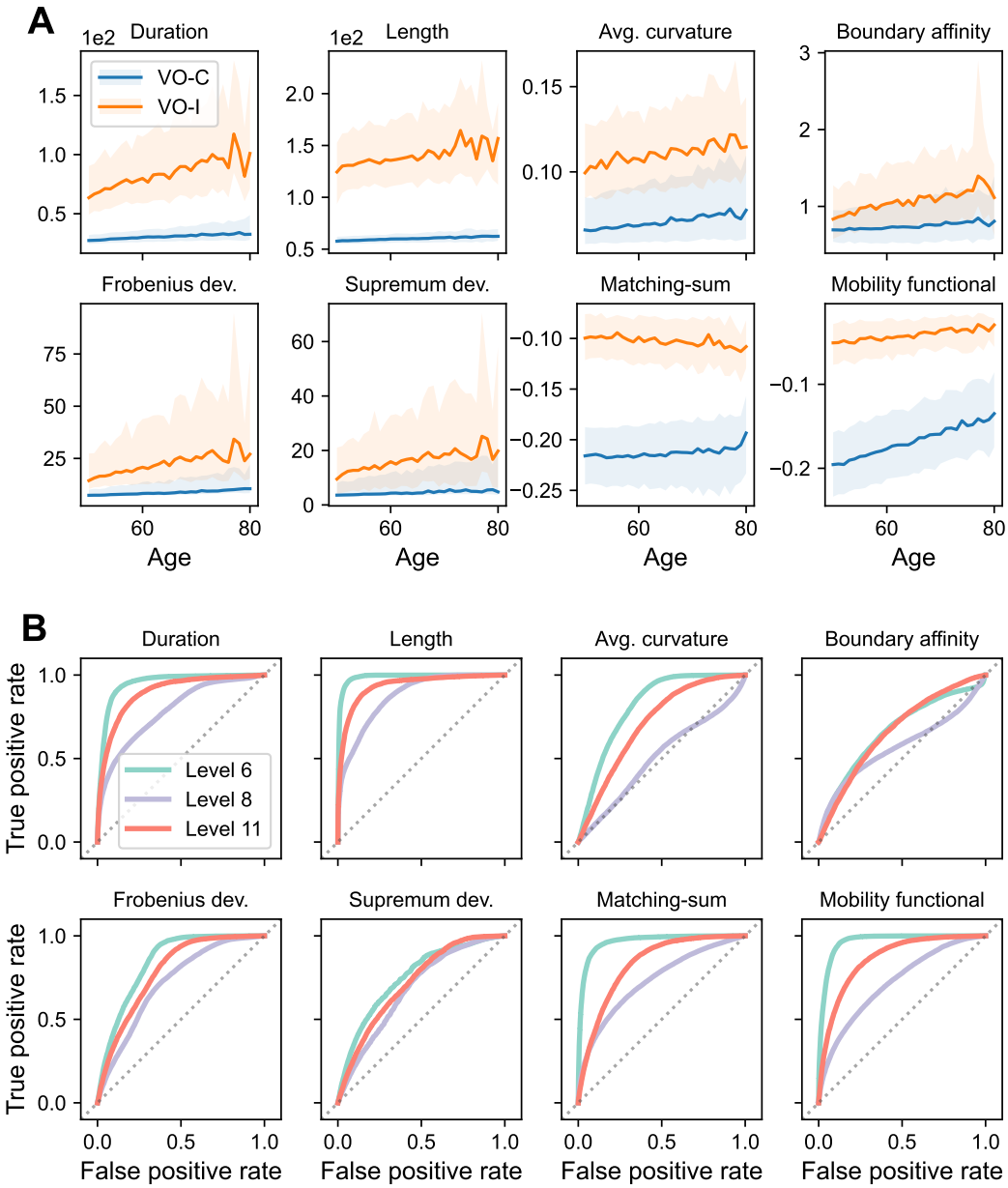


Figure 6.2: **A**. Median and inter-quartile spread of the measures across age for Level 6 (males and females are grouped together). The two groups correspond to the correct visiting order (blue) and the incorrect visiting order (orange). **B**. Receiver operating characteristic (ROC) curves for the visiting order.

measures when plotted with respect to age and separated by VOC. This figure was generated using the data for Level 6, but similar results can be seen for Levels 8 and 11 in the Appendix E, Figures E.1 and E.2.

For all eight metrics represented in this figure, the median of the group with correct VO (in blue) always lies beneath the median for the group with incorrect VO (in orange). This intuitively makes sense: on average, we can expect that a trajectory will have to be longer, make more turns and therefore have a higher curvature, or visit less navigated areas of the map if the player makes a mistake while visiting the checkpoints.

Perhaps more interesting is the amount to which the inter-quartile shaded regions overlap. We can see the we have zero overlap for Duration, Length, Matching-sum, and the Mobility functional, while we do have an overlap for the other measures. This separation is desirable because it is related to how easily we can tell the groups apart — more on this below.

It is interesting that we can see the effect that age has in two ways: in an overall trend for the measures to increase with age, and in a larger interquartile spread (particularly for the orange group) which indicates a larger number of outlier and more variability. The two measures that do not follow this are the Matching-sum and the Mobility functional which are bounded above by construction, so that even the most pathological trajectory does not become a very big outlier. We also note that the Matching-sum measure remains relatively flat with respect to age. This behaviour is what we were looking for when we proposed the relative measures because we would like the moving window to compensate for the difference in navigation ability with respect to age, but we found it hard to achieve.

Figure 6.2B has the receiver operating characteristic (ROC) curves for the eight metrics and the three levels studied (in different colours). The curves were computed with a moving threshold so that every score underneath it is considered a positive (correct visiting order) example. In this experiment we mix all of the players irrespective of age, so that the results reflect an average for the normative players aged 50 to 80 years. A curve which is above the diagonal indicates that we can separate the two groups better than random guessing, and this is the case for 22 of the 24 curves. We have a clear ranking with Level 6 always coming at the top and Level 8 coming last. The area under the curve (AUC) scores are reported in Table 6.4.

The two curves which do not stay consistently above the diagonal were obtained for Level 8, for the average curvature (AUC F: 0.475 and M: 0.471) and the boundary affinity measure (AUC F: 0.628 and M: 0.493). As we have explained before, Level 8 is where we get a significantly lower percentage of players completing the level correctly and we consider it the hardest level. The two groups can be told apart with other measures but not with curvature and affinity. According to our medical colleagues, at-risk patients do show a tendency to navigate closer to the boundary of levels and we hope to carry a proper calibration of the parameters r_{in} and r_{out} that can help us better show the separation between groups using the boundary affinity metric. As for the curvature, our hypothesis is that the averaging that we do might be obscuring the trajectories that are incredibly windy only in some sections and not in others.

The AUC scores allow us to positively conclude that the overall separation between the VOC groups is most clearly reflected by measuring the length (mean AUC across the levels

	Level 6		Level 8		Level 11		Mean
	F	M	F	M	F	M	
Duration	0.946	0.951	0.781	0.782	0.883	0.903	0.874
Length	0.986	0.991	0.872	0.871	0.925	0.941	**0.931**
Avg. curv.	0.726	0.799	0.475	0.471	0.582	0.652	0.617
Bdy. Affty.	0.656	0.553	0.628	0.493	0.611	0.615	0.593
Frob. dev.	0.823	0.856	0.690	0.734	0.750	0.816	0.778
Sup. dev.	0.742	0.745	0.650	0.690	0.681	0.733	0.707
Match. sum	0.960	0.968	0.725	0.728	0.842	0.839	0.844
Mob. funct.	0.956	0.974	0.696	0.726	0.845	0.872	0.845

Table 6.4: Area under the curve (AUC) for the receiver operating characteristic (ROC) curves of the proposed measures. In Figure 6.2 we grouped the two genders for clarity but here we report the results for each gender separately.

and genders: 0.931). The other measures that do comparatively very well are the duration of the attempt and the Matching-sum and mobility functional measures. These results serve as validation for the analysis that Coughlan et al. carried out [18]: length and duration, arguably the simplest of all the measures discussed, have a high power of discrimination between the most obviously unsuccessful attempts.

6.4 Performance of the clinical cohorts

6.4.1 Visiting order

From the analysis presented in Section 6.3, we know that the order in which the checkpoints are visited can be used as a simple definition of what we consider to be a problematic trajectory.

We compute the visiting order correctness for the clinical groups and give the results in Table 6.5. According to this crude measure, the $\epsilon 3$ homozygous group has the least trouble finding their way: all of the patients in the group complete Level 6 in the right order and all but one patient complete Level 11 correctly. Level 8 has a failure rate just under 50%, which we take as normal and actually slightly higher than normal because this is better than the rate among the normative dataset (see Table 6.3).

We compare these numbers with the $\epsilon 4$ carriers and we find a slightly deteriorated performance for the group as a whole. Roughly 80% of these patients complete Level 6 in the correct order which is now closer to what we observe with the normative population. The same is true for Level 8, while Level 11 is completed corrected at a slightly higher rate than expected.

It is very interesting to find that the $\epsilon 4$ carriers behave more like the normative population than the non-carriers. We do not necessarily expect this because we expect the normative population to be representative of the population as a whole, which would make the non-carriers a majority of the dataset due to the allele frequencies. We hypothesise that the participants of the study are paying more attention to the tasks because they are aware that they are part of a study — they are probably in a laboratory — and this heightens not necessarily their performance but at least the crude measure of visiting the checkpoints in order, which can probably be be

Group	n	Level 6				Level 8				Level 11			
		VO-C	%	VO-I	%	VO-C	%	VO-I	%	VO-C	%	VO-I	%
AD	18	7	39	11	61	1	6	17	94	8	44	10	56
e3e3	28	28	100	0	0	15	54	13	46	27	96	1	4
e3e4	31	26	84	5	16	13	42	18	58	28	90	3	10
e4e4	5	5	100	0	0	5	100	0	0	3	60	2	40

Table 6.5: Number and fraction of trajectories in the clinical groups that complete each level visiting the checkpoints in the correct order.

done by paying more attention. On the other hand, the home users that compose the normative dataset probably have a very low stake in the study and this will likely impact their attention.

The last group that we want to mention is the group of patients diagnosed with AD. For Levels 6 and 11 this group has a failure rate approximately twice as high as the normative population, and for Level 8, the hardest of them, only a single patient completes the task visiting the checkpoints in the correct order giving us a 94% failure rate, by far the worst performance for any group and level.

After this high-level analysis, we now compare the rest of the measures in more detail using the scores that each individual obtains.

6.4.2 Percentile scores

In their study of at-genetic-risk patients, Coughlan et al. briefly benchmark the trajectory lengths of the $\epsilon 4$ carriers inside a series of plots that show the score of the patients on top of a histogram for the sub-population of SHQ players with matched age, gender and education level [18].

Here we formalise this idea. For each of the measures introduced previously, we compute the percentile of the score obtained by a clinical patient relative to the normative population. The percentile and not the raw value of a metric has the advantage that it is less affected by outliers and it has a direct and intuitive interpretation: a patient that scores within the 95-th percentile did worse than all but 5% of the normative population, matched for age and gender.

Note that the comparison on which the calculation of the percentile score is based is similar to the comparison that defines the relative measures. Take, for example, the mobility functional $\Phi[\boldsymbol{\gamma}, \{\Gamma_d\}_{d=a-5}^{a+5}]$ where a is the age of the patient. To compute the percentile, we must fist calculate $\Phi[\boldsymbol{\gamma}', \{\Gamma_d\}_{d=a-5}^{a+5}]$ $\forall \boldsymbol{\gamma}' \in \Gamma_a$ and then compare the rank of the patient inside this set. The purpose of the moving window in the relative measures is to correct for declining ability with respect to age and we can ask if we succeeded in doing so; however, the percentile score is the ultimate age correction because everyone is measured by comparing with the normative population with the exact same age.

Preliminary results

For each of the three levels studied, we compute the percentiles of the scores obtained by the genotyped patients and the patients diagnosed with AD.

As we discuss below, the patients diagnosed with AD had the most trouble navigating and

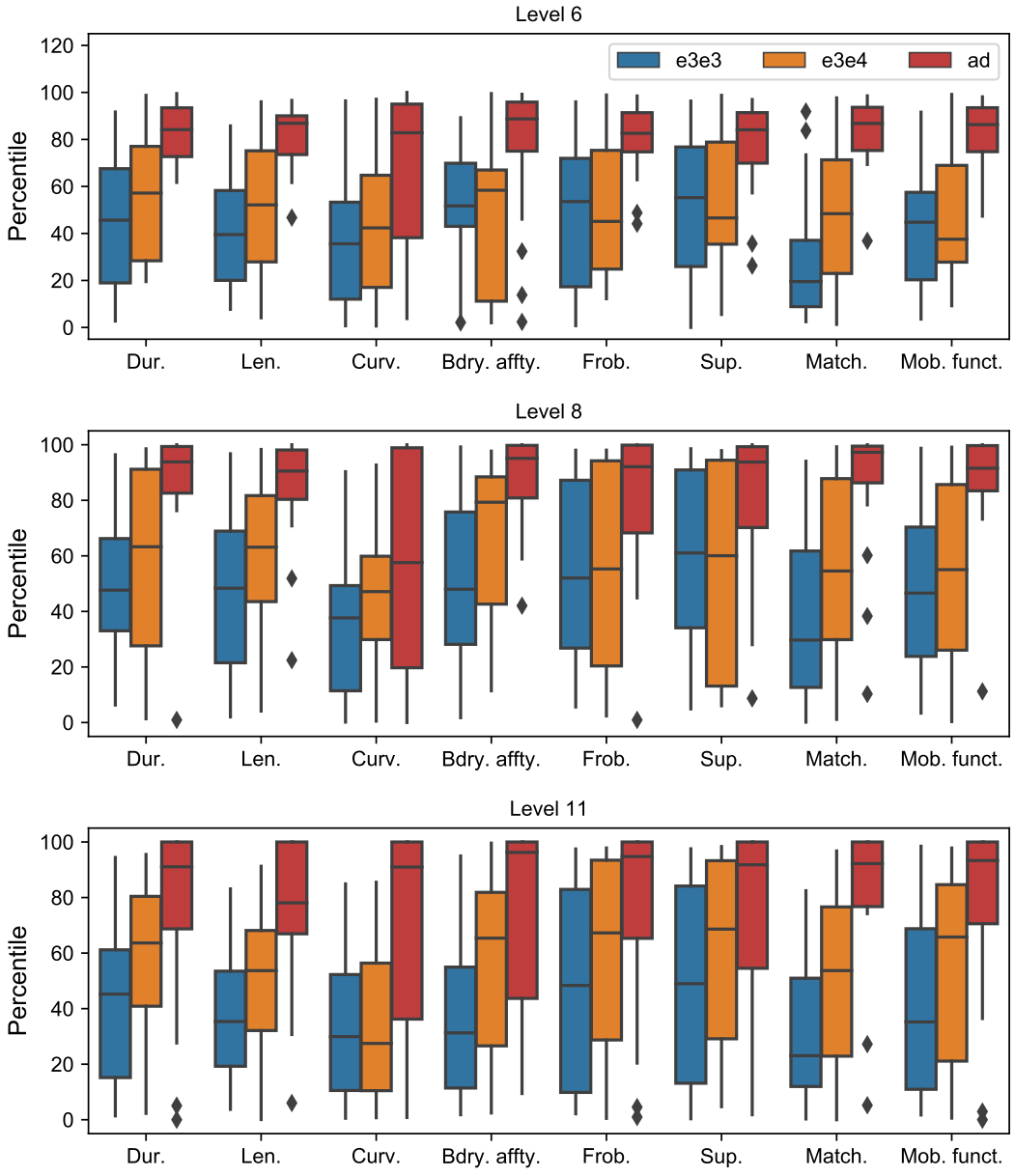


Figure 6.3: Comparison of the percentiles scores for the different genetic groups and the cohort of patients with AD, across levels and for each metric (higher scores represent a worse performance). The median is indicated with a black horizontal line and the box gives the interquartile spread.

this is reflected by a higher length, higher, duration, etc. However, some of the participants in the AD study were in so much distress that they stopped playing the Sea Hero Quest game altogether without attempting the three wayfinding levels analysed here: from the original 21 patients diagnosed with AD, 18 of them attempted Level 6; 15 of them attempted Level 8; and only 11 attempted Level 11. Obviously, the measures are missing for those players, and this could have the effect of biasing the average scores for the group as a whole because only those individuals that do not struggle too much remain. To mitigate for this, we assign the percentile of 100 to patients that did not complete Levels 8 and 11 (in this we take the 18 patients that finished level 6 as the full group; we no longer consider the three patients that did not get to complete the first wayfinding level).

We compare the percentiles for the eight different metrics Figure 6.3. Overall, we observe a general trend in which the $\epsilon 3$ homozygous patients score lower than the $\epsilon 4$ carriers (indicating better performance) and these in turn score lower than patients with an AD diagnosis. However, there are exceptions to this trend depending on the level and the measure chosen.

For Level 6, the median score of the $\epsilon 4$ carriers is lower than the homozygous $\epsilon 3$ patients according to the Frobenius, supremum and mobility functional measures. Interestingly, this is not true for the Matching-sum measure. Though these four metrics are based on a comparison with a normative OD matrix, the results can vary significantly and impact the average score of the groups. This is probably due to the fact that the matrix norms, though normalised, are still sensitive to the large entries of the normative OD matrix. Furthermore, it is no coincidence that Level 6 seems to be the most impacted because it the most open and there are many more regions of the map that can be visited. For Level 11 which is narrower, the four relative measures agree that the $\epsilon 4$ group is performing worse than the $\epsilon 3$ group.

An interesting measure from the point of view of AD patients is average curvature. We find a significantly larger red box than for the other measures, indicating a greater inter-quartile spread and that many of the attempts made by the AD group are not considered outliers from the point of view of their windy trajectories. This is due to the fact that both the patients with dementia and the players in the normative dataset struggle to steer the boat in a straight line, probably because they are all between 50 and 80 years old and not very familiar with playing games on a mobile device.

Before we draw any more conclusions about the comparison of the proposed measures, let us fist modify our computation of the percentiles by making use of the visiting order.

Improvement to the results using VOC

We can make an amendment to the computation of the percentiles scores presented previously: let us take, for the comparison, the subset of the normative population with matching age and gender but also with a correct visiting order. When doing this, an important question to ask ourselves is the following: if we exclude the trajectories that exhibit a wrong visiting order, to what degree are we also excluding the populations more at risk, or even the individuals with pathologies?

Obviously, we have no way of knowing whether the players around the world are healthy or

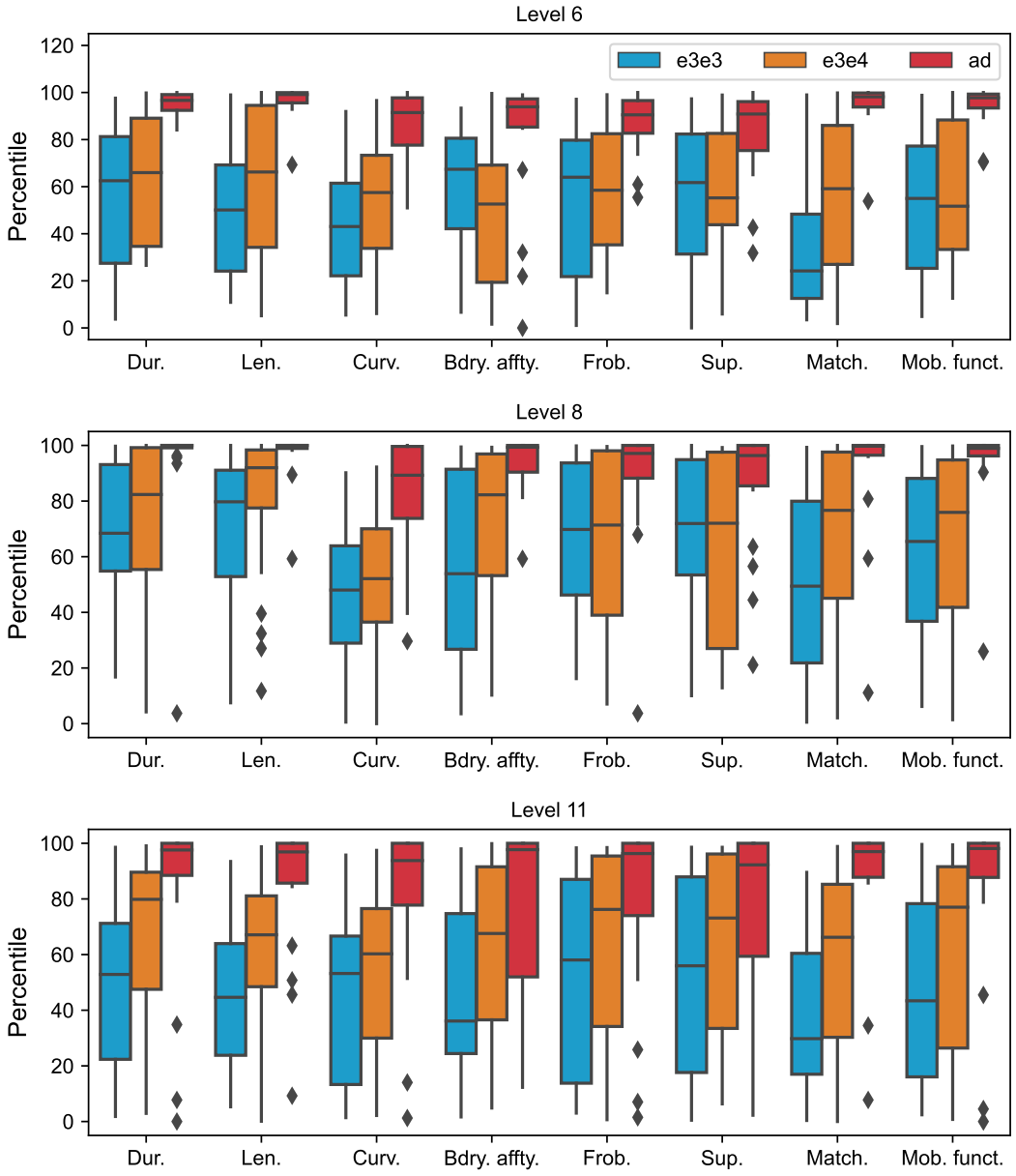


Figure 6.4: Comparison of the percentiles scores for the different genetic groups and the cohort of patients with AD, across levels and for each metric (higher scores represent a worse performance). Compared to Figure 6.3, the difference is that we calculate the percentiles by doing a comparison only against the subset of the normative population that has a correct visiting order.

if they have been diagnosed with AD or MCI. It is also impossible to know who is an $\epsilon 4$ carrier and is therefore at increased genetic risk of developing the disease. It is therefore impossible to answer the previous question with the Sea Hero Quest database as it currently stands. However, we can hope that the pool of people that played the game is large enough to be representative of the general population in terms of allele frequency and disease prevalence.

The percentiles resulting when we exclude the incorrect visiting orders are shown in Figure 6.4. Compared to what we had previously, we can easily see that the dementia patients are considered outliers in almost all the cases. Average curvature is still an exception but this is to be expected since the explanation that we gave before still holds (the normative trajectories with the correct visiting order are still not made of straight lines).

6.5 A proposal for a Bayesian formulation of diagnosis

The analysis shown previously tells us that patients diagnosed with dementia tend to score very high in several measures when compared with the normative population (measured in terms of a percentile score). Equally, we have seen that patients in one genetic group tend to have higher scores that patients in the other group. Given all of this, it is natural to ask whether this knowledge has any diagnostic power.

Suppose that we are trying to determine the risk of AD of an individual that has not had a diagnosis. For the time being, let us also suppose that the version of the APOE gene of the individual is unknown; what little information we have is the individual's age a and gender g. The individual plays the SHQ game and attempts the wayfinding levels. For a particular level, suppose that we find that the percentile they obtain for one of the measures introduced is p, when compared to the normative dataset.

One way to quantify the risk is by measuring the plausibility of the hypothesis "the individual has AD" given their percentile p and their demographic characteristics. In a Bayesian setting, this is measured using the posterior probability

$$\mathbb{P}(\mathrm{AD} \mid p, a, g) \propto \mathbb{P}(p \mid \mathrm{AD}, a, g)\, \mathbb{P}(\mathrm{AD}, a, g)\,. \tag{6.14}$$

In this equations, the prior $\mathbb{P}(\mathrm{AD}, a, g)$ can be estimated using the AD prevalence rates which are published in the UK by the Office for National Statistics (see Table 6.6). The likelihood term $\mathbb{P}(p \mid \mathrm{AD}, a, g)$ can be estimated from the previous analysis. One way to do this is to find the level and measure which are being looked at in Figure 6.4, and determining the fraction of the AD patients which fall above or below the percentile p. Obviously, the sample size $n = 18$ which we have access to in too small for a reliable estimation, but this is data that could be acquired in a larger study.

The term we omitted from Equation 6.14 is the denominator $\mathbb{P}(p)$ in the right hand side, called the evidence. We could argue that the most natural way to compute is $\mathbb{P}(p) = p$, which makes use of the fact that p is a percentile. However, we can get around using it entirely if

Age bracket	Females (%)	Males (%)	Total (%)
60-64	0.9	0.9	0.9
65-69	1.8	1.5	1.7
70-74	3.0	3.1	3.0
75-79	6.6	5.3	6.0
80-84	11.7	10.3	11.1
85-89	20.2	15.1	18.3
90-94	33.0	22.6	29.9
95+	44.2	28.8	41.1

Table 6.6: The prevalence rates of Alzheimer's Disease in the United Kingdom for the year 2014. Published by the Office for National Statistics.

instead we compute the odds

$$\mathbb{O}(\mathrm{AD} \mid p, a, g) := \frac{\mathbb{P}(\mathrm{AD} \mid p, a, g)}{\mathbb{P}(\overline{\mathrm{AD}} \mid p, a, g)}\,, \tag{6.15}$$

$$= \frac{\mathbb{P}(p \mid \mathrm{AD}, a, g)}{\mathbb{P}(p \mid \overline{\mathrm{AD}}, a, g)}\,\mathbb{O}(\mathrm{AD}, a, g)\,, \tag{6.16}$$

where $\overline{\mathrm{AD}}$ is the hypothesis "the individual *does not* have AD", and the prior odds are

$$\mathbb{O}(\mathrm{AD}, a, g) = \frac{\mathbb{P}(\mathrm{AD}, a, g)}{1 - \mathbb{P}(\mathrm{AD}, a, g)}\,. \tag{6.17}$$

The one term that remains to define or to say how we can find it is the likelihood term $\mathbb{P}(p \mid \overline{\mathrm{AD}}, a, g)$. At the moment, the best we can do it is to take the cohort of genotyped patients because we know that these individuals are healthy (though maybe at risk), mix the groups irrespective of the allele they carry and compute the fraction that score above or below the percentile p. However, the group sizes in this cohort of patients is not representative of the population as a whole (given that we have almost as many $\epsilon4$ carriers as $\epsilon3$ homozygotes) and it is also completely lacking the $\epsilon2$ carriers. Thus, this approach remains far from the ideal solution. A better approach would be to collect new data in a bigger study, and have patients diagnosed as healthy play the SHQ game the same way the patients diagnosed with AD do.

6.5.1 Genotype inference

In a similar way, we can compute the plausibility of the hypothesis "the individual is part of genetic group X" for $X \in \{\epsilon3\epsilon3, \epsilon3\epsilon4\}$ (or new genetic groups if they become available). In our calculation we would use the allele frequencies (for example those in Table 6.1) to compute the priors, and we would find the likelihood term again by counting the number of scores above and below the percentile p for the genetic group X. In this case, we would also benefit from larger numbers in each group. Note that this inference procedure can be done in addition to the AD diagnosis described above because the genetic group is independent to the fact that we can determine the plausibility of the individual suffering from Alzheimer's disease.

It was recently brought to our attention that the goal of our Bayesian proposal is similar

to the more standard regression setting, for example with Generalised Linear Models (GLMs) in general or Logistic Regression in particular. The added benefit in the Bayesian formulation of being able to do sequential updates is probably not crucial for the application that we are proposing. Instead, it makes sense to use a regression model that controls for the demographics, and which can easily be extended to control for other variables that we have not used such as education or handedness. A colleague is working on precisely this extension, and when included, we will submit the manuscript of the article introducing our proposed new metrics (see the end of Section 1.6).

Chapter 7

Conclusions

The two overarching themes in this thesis have been the use of networks as a modelling tool and the analysis of spatial data. A big portion of our work has been motivated by our collaboration with Dunnhumby, a company which specialises in customer data science and wanted to employ networks to study the relationship between retail customers and the stores they shop in. Yet, we tried to develop general methods that can also be applied elsewhere. For example, we collaborated with medical colleagues that had an interesting problem to solve about a dataset used in research for Alzheimer's disease, and we found that some of the same network tools were useful.

Overall, we deal with humans and the way in which humans move. We find it fascinating that there are still so many open questions in the field, and our goal has been to shine the light of Network Science on the problems which we found to be amenable.

In Chapter 2, we presented our first model to analyse a year's worth of anonymised retail data (aggregated temporally) using a weighted bipartite network. We applied a community detection algorithm — the map equation — and we found groups of both stores and customers that were clustered geographically. This result arguably does not say a great deal that is new, given that we can expect the strongest interactions to take place amongst customers and the stores that are easily within their reach so we set out to develop a way to find the "space-corrected" mesoscale structure of a network. This is the topic of the following two chapters.

In Chapter 3 we gave a treatment of two popular human mobility models, the gravity and the radiation model, and we reinterpreted them in a statistical way by means of a constrained formulation. This is important because in Chapter 4 we treated the prediction of the models as random variables to define what we call the spatial backbone of a network. This newly defined graph holds the edges that are significant because their associated weight is either higher or lower than expected. We can fit a degree-corrected stochastic block model to the backbone to find groups of nodes that interact particularly strongly (or particularly weakly) compared to a baseline that we are free to choose. We employed this methodology to study a unipartite projection of the customer-store network that models the interaction between the stores, as well as the network of commuting flows from the 2011 UK census.

In Chapter 5 we revisit the anonymised retail dataset — now covering four years of records —

from the perspective of the customers (and to a lesser extent the stores). We define fifteen metrics that can be computed for each customer using a collection of baskets, and we investigate the seasonal variation of shopping habits as well as the shock created by the COVID-19 pandemic. We also propose a novel way to reconstruct a possible transfer of demand, and we apply it inside London and at then a national level. Another key aspect that we study is the mobility of the customers, particularly the radius of gyration that measures the linear size occupied by their trajectories. Again, we can see different habits, and in particular we find a strong signal for a group of customers that has a zero mean displacement during the summer holidays.

We changed the application in Chapter 6 and we worked with navigation inside a virtual game that was developed to help in the early detection of Alzheimer's Disease. We had access to two main sources of data, the attempts made by two clinical cohorts that have either been diagnosed with dementia or had their genotype sequenced, and also a large population of volunteers that serves as a normative benchmark. We defined new measures of navigational ability which encode geometric information (such as curvature) or the similarity with the trajectories found in amongst the normative population.

Working with real-world data is always a source of real-world problems and we have to acknowledge that it imposed some limitations on the research that we were able to do. In addition to the anonymised shopping records that we had access to, we would have liked to be able to locate the households even approximately. The *Retailer* knows the home address that the customers used to sign-up to the loyalty card programme, but due to privacy concerns it was always ruled out to grant us permission to access this. This meant that we were never able to include a household's distance to the different stores in our modelling.

In terms of the modelling decisions that we made, we know that the as-the-crow-flies distance that we employed in chapters 3 to 5 is a simple choice that can be considered too rough an approximation because it completely ignores the underlying topography and there are alternatives that we could have used instead. The flow reconstruction analysis inside London would benefit greatly from using the distance on the road network because the fact that the Thames puts a physical barrier between parts of the city is currently not being accounted for.

The method that we propose for fitting the stochastic block model to spatial networks currently places the same importance on all the statistically significant edges. While this is a desired feature that preserves as much information as possible when we sparsify the original network and we build the spatial backbone, it comes with the drawback that mobility flows vary over several orders of magnitude and considering them all on an equal footing dilutes the importance of the largest entries — flows that are as high as 10^4 for the commuting network and 10^5 for the retail network — which, it can be argued, should receive special attention. More work needs to be done to explore the effect of varying the significance threshold which can be a successful way forward to mitigate this behaviour.

With more time, we would be curious to study in more detail the spatial backbone, a signed network. In particular, we would like to know if the datasets studied here lead to structurally balanced networks, and consider more deeply the implications of a positive or negative answer.

In Chapter 5 we claim to see the effect of households travelling during the summer months

but we had no way of validating this experimentally. However, this could be ascertained by Dunnhumby using surveys or possibly alternative data sources that they already hold.

An obvious extension that would benefit different analyses that have been presented in our research, would be employing other mobility models besides the well-known gravity and radiation models. These two models are well suited for flows between locations (and they are appropriate for our chosen applications of the retail and the commuting networks) but there are other spatial systems in which the interactions do not necessarily take the form of flows. One example is the transitions between pixels that we encountered in Sea Hero Quest. We were able to successfully define an origin-destination matrix, but it would probably be a stretch to describe the system further in terms of flows. This is why models of a different kind might still be needed to incorporate the effect of space and spatial constraints.

As for the new metrics introduced to quantify spatial navigation ability, we have work to do to calibrate the parameters in the boundary affinity metric. We would also like to develop a methodology to use our metrics "online", meaning that we have a live assessment of how uncommon is a trajectory as it is being travelled. According to our medical colleagues, this could be an exciting application for our work because in their field they mostly rely on length and duration. Even if we have seen the two metrics to have a good discriminative power, the assessment needs to be over to be able to calculate them.

All in all, we have made the case that we can apply tools and modelling ideas from Network Science to study spatial systems. Our research focused on a few applications for which we had the data, but thinking about the ever-increasing need to analyse the spatial information that is collected with our digital gadgets, we are hopeful that we are contributing useful ideas that can be applied elsewhere.

Bibliography

[1] Emmanuel Abbe. 'Community Detection and Stochastic Block Models: Recent Developments'. In: *Journal of Machine Learning Research* 18.1 (Apr. 2018), p. 86.

[2] Carl S. Adorf et al. 'Simple data and workflow management with the signac framework'. In: *Computational Materials Science* 146.C (2018), pp. 220–229. DOI: 10.1016/j.commatsci.2018.01.035.

[3] A. Arenas et al. 'Size reduction of complex networks preserving modularity'. In: *New Journal of Physics* 9.176 (June 2007). ISSN: 1367-2630. DOI: 10.1088/1367-2630/9/6/176.

[4] Ulf Aslak, Peter Møllgaard and Laura Alessandretti. *Travel in Denmark*. Sept. 2020. URL: https://covid19.compute.dtu.dk/visualizations/telco_brush/ (visited on 25/07/2022).

[5] Ulf Aslak and Peter Møllgard. *Distance Traveled*. July 2020. URL: https://covid19.compute.dtu.dk/visualizations/how_people_move_distance_traveled/ (visited on 25/07/2022).

[6] James P. Bagrow. 'Communities and bottlenecks: Trees and treelike networks have high modularity'. In: *Physical Review E* 85.6 (June 2012). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.85.066118.

[7] D. Balcan et al. 'Multiscale mobility networks and the spatial spreading of infectious diseases'. In: *Proceedings of the National Academy of Sciences* 106.51 (Dec. 2009), pp. 21484–21489. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0906910106.

[8] Michael J. Barber. 'Modularity and community detection in bipartite networks'. In: *Physical Review E* 76.6 (Dec. 2007). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.76.066102.

[9] Hugo Barbosa et al. 'Human mobility: Models and applications'. In: *Physics Reports* 734 (Mar. 2018), pp. 1–74. ISSN: 03701573. DOI: 10.1016/j.physrep.2018.01.001.

[10] Marc Barthelémy. *Mophogenesis of Spatial Networks*. Lecture Notes in Morphogenesis. Springer, 2018. ISBN: 978-3-319-20564-9.

[11] Marc Barthélemy. 'Spatial networks'. In: *Physics Reports* 499 (Feb. 2011), pp. 1–101. ISSN: 03701573. DOI: 10.1016/j.physrep.2010.11.002.

[12] Vincent D Blondel et al. 'Fast unfolding of communities in large networks'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/P10008.

[13] Richard A Brualdi, Seymour V Parter and Hans Schneider. 'The diagonal equivalence of a nonnegative matrix to a stochastic matrix'. In: *Journal of Mathematical Analysis and Applications* 16.1 (Oct. 1966), pp. 31–50. ISSN: 0022247X. DOI: 10.1016/0022-247X(66)90184-3.

[14] Federica Cerina et al. 'Spatial Correlations in Attribute Communities'. In: *PLoS ONE* 7.5 (May 2012). Ed. by Sergio Gómez, e37507. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0037507.

[15] Serina Chang et al. 'Mobility network models of COVID-19 explain inequities and inform reopening'. In: *Nature* 589.7840 (), pp. 82–87. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-2923-3.

[16] Aaron Clauset, M. E. J. Newman and Cristopher Moore. 'Finding community structure in very large networks'. In: *Physical Review E* 70.6 (Dec. 2004). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.70.066111.

[17] Gillian Coughlan et al. 'Spatial navigation deficits — overlooked cognitive marker for preclinical Alzheimer disease?' In: *Nature Reviews Neurology* 14.8 (Aug. 2018), pp. 496–506. ISSN: 1759-4758, 1759-4766. DOI: 10.1038/s41582-018-0031-x.

[18] Gillian Coughlan et al. 'Toward personalized cognitive diagnostics of at-genetic-risk Alzheimer's disease'. In: *Proceedings of the National Academy of Sciences* 116.19 (May 2019), pp. 9285–9292. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1901600116.

[19] A. Coutrot et al. 'Entropy of city street networks linked to future spatial navigation ability'. In: *Nature* 604.7904 (Apr. 2022), pp. 104–110. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-022-04486-7.

[20] Antoine Coutrot et al. 'Global Determinants of Navigation Ability'. In: *Current Biology* 28.17 (Sept. 2018), 2861–2866.e4. ISSN: 09609822. DOI: 10.1016/j.cub.2018.06.009.

[21] *Covid-19 Impact Monitor reveals UK population moves drops by 98%*. Apr. 2020. URL: https://www.ox.ac.uk/news/2020-04-15-covid-19-impact-monitor-reveals-uk-population-moves-drops-98 (visited on 25/07/2022).

[22] Peng-Bi Cui, Francesca Colaiori and Claudio Castellano. 'Effect of network clustering on mutually cooperative coinfections'. In: *Physical Review E* 99.2 (Feb. 2019). ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.99.022301.

[23] Leon Danon et al. 'Comparing community structure identification'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2005.09 (Sept. 2005), P09008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2005/09/P09008.

[24] Jean Charles Delvenne et al. 'The stability of a graph partition: A dynamics-based framework for community detection'. In: *Modeling and Simulation in Science, Engineering and Technology* 55 (2013), pp. 221–242. ISSN: 978-1-4614-6729-8. DOI: 10.1007/978-1-4614-6729-8_9.

[25] A. P. Dempster, N. M. Laird and D. B. Rubin. 'Maximum Likelihood from Incomplete Data Via the EM Algorithm'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* (Sept. 1977). DOI: 10.1111/j.2517-6161.1977.tb01600.x.

[26] Xiaowen Dong et al. 'Segregated interactions in urban and online space'. In: *EPJ Data Science* 9.1 (Dec. 2020), p. 20. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-020-00238-7.

[27] P. Expert et al. 'Uncovering space-independent communities in spatial networks'. In: *Proceedings of the National Academy of Sciences* 108.19 (May 2011), pp. 7663–7668. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1018962108.

[28] Stephen E. Fienberg. 'An Iterative Procedure For Estimation in Contingency Tables'. In: *The Annals of Mathematical Statistics* 41.3 (1970), pp. 907–917.

[29] Rémi Flamary et al. 'POT: Python Optimal Transport'. In: *Journal of Machine Learning Research* 22.78 (2021), pp. 1–8.

[30] S. Fortunato and M. Barthélemy. 'Resolution limit in community detection'. In: *Proceedings of the National Academy of Sciences* 104.1 (Jan. 2007), pp. 36–41. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0605965104.

[31] Santo Fortunato. 'Community detection in graphs'. In: *Physics Reports* 486.3-5 (2010), pp. 75–174. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2009.11.002.

[32] Santo Fortunato and Darko Hric. 'Community detection in networks: A user guide'. In: *Physics Reports* 659 (Nov. 2016), pp. 1–44. ISSN: 03701573. DOI: 10.1016/j.physrep.2016.09.002.

[33] Bailey K. Fosdick et al. 'Configuring Random Graph Models with Fixed Degree Sequences'. In: *SIAM Review* 60.2 (Jan. 2018), pp. 315–355. ISSN: 0036-1445, 1095-7200. DOI: 10.1137/16M1087175.

[34] A. Stewart Fotheringham. *Spatial Interaction Models*. 2001.

[35] A. Stewart Fotheringham, Michael Batty and Paul A. Longley. 'Diffusion-limited aggregation and the fractal nature of urban growth'. In: *Papers of the Regional Science Association* 67.1 (Dec. 1989), pp. 55–69. ISSN: 1056-8190, 1435-5957. DOI: 10.1007/BF01934667.

[36] Oliver Gardiner and Xiaowen Dong. 'Mobility Networks for Predicting Gentrification'. In: *Complex Networks & Their Applications IX*. Vol. 944. Series Title: Studies in Computational Intelligence. Springer International Publishing, 2021, pp. 181–192. ISBN: 978-3-030-65350-7 978-3-030-65351-4. DOI: 10.1007/978-3-030-65351-4_15.

[37] Sean Gillies et al. *Shapely: manipulation and analysis of geometric objects*. toblerity.org, 2007–. URL: https://github.com/Toblerity/Shapely.

[38] M. Girvan and M. E. J. Newman. 'Community structure in social and biological networks'. In: *Proceedings of the National Academy of Sciences* 99.12 (June 2002), pp. 7821–7826. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.122653799.

[39] Marta C. González, César A. Hidalgo and Albert-László Barabási. 'Understanding individual human mobility patterns'. In: *Nature* 453.7196 (June 2008), pp. 779–782. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature06958.

[40] Benjamin H. Good, Yves-Alexandre de Montjoye and Aaron Clauset. 'Performance of modularity maximization in practical contexts'. In: *Physical Review E* 81.4 (Apr. 2010). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.81.046106.

[41] *Google COVID-19 Community Mobility Report*. 2020. URL: https://www.google.com/covid19/mobility?hl=en (visited on 25/07/2022).

[42] Aditya Grover and Jure Leskovec. 'node2vec: Scalable feature learning for networks'. In: *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. 2016, pp. 855–864.

[43] R. Guimerà et al. 'The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles'. In: *Proceedings of the National Academy of Sciences* 102.22 (May 2005), pp. 7794–7799. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0407994102.

[44] Roger Guimerà, Marta Sales-Pardo and Luís A. Nunes Amaral. 'Modularity from fluctuations in random graphs and complex networks'. In: *Physical Review E* 70.2 (Aug. 2004). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.70.025101.

[45] Roger Guimerà, Marta Sales-Pardo and Luís A. Nunes Amaral. 'Module identification in bipartite and directed networks'. In: *Physical Review E* 76.3 (Sept. 2007). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.76.036102.

[46] Furkan Gursoy and Bertan Badur. 'Extracting the signed backbone of intrinsically dense weighted networks'. In: *Journal of Complex Networks* 9.5 (Sept. 2021). Ed. by Ernesto Estrada, cnab019. ISSN: 2051-1310, 2051-1329. DOI: 10.1093/comnet/cnab019.

[47] Paul W. Holland, Kathryn Blackmond Laskey and Samuel Leinhardt. 'Stochastic blockmodels: First steps'. In: *Social Networks* 5.2 (June 1983), pp. 109–137. ISSN: 03788733. DOI: 10.1016/0378-8733(83)90021-7.

[48] Lawrence J. Hubert. 'Some applications of graph theory to clustering'. In: *Psychometrika* 39.3 (Sept. 1974), pp. 283–309. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/BF02291704.

[49] Jonathan Q. Jiang. 'Stochastic block model and exploratory analysis in signed networks'. In: *Physical Review E* 91.6 (June 2015), p. 062805. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.91.062805.

[50] Pablo Kaluza et al. 'The complex network of global cargo ship movements'. In: *Journal of The Royal Society Interface* 7.48 (July 2010), pp. 1093–1103. ISSN: 1742-5689, 1742-5662. DOI: 10.1098/rsif.2009.0495.

[51] Brian Karrer and M. E. J. Newman. 'Stochastic blockmodels and community structure in networks'. In: *Physical Review E* 83.1 (Jan. 2011). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.83.016107.

[52] B. W. Kernighan and S. Lin. 'An efficient heuristic procedure for partitioning graphs'. In: *The Bell System Technical Journal* 49.2 (Feb. 1970), pp. 291–307. ISSN: 0005-8580. DOI: 10.1002/j.1538-7305.1970.tb01770.x.

[53] M. Kivelä et al. 'Multilayer networks'. In: *Journal of Complex Networks* 2.3 (Sept. 2014), pp. 203–271. ISSN: 2051-1310, 2051-1329. DOI: 10.1093/comnet/cnu016.

[54] Gautier Krings et al. 'Urban gravity: a model for inter-city telecommunication flows'. In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.07 (July 2009), p. L07003. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2009/07/L07003.

[55] R. Lambiotte, J.-C. Delvenne and M. Barahona. 'Laplacian Dynamics and Multiscale Modular Structure in Networks'. In: *IEEE Transactions on Network Science and Engineering* 1.2 (July 2014), pp. 76–90. ISSN: 2327-4697. DOI: 10.1109/TNSE.2015.2391998.

[56] Daniel B. Larremore et al. 'Test sensitivity is secondary to frequency and turnaround time for COVID-19 screening'. In: *Science Advances* 7.1 (Jan. 2021). ISSN: 2375-2548. DOI: 10.1126/sciadv.abd5393.

[57] Sune Lehmann and Ulf Aslak. *Population landscape*. May 2022. URL: https://covid19.compute.dtu.dk/posts/population_landscape/ (visited on 25/07/2022).

[58] Maxime Lenormand, Aleix Bassolas and José J. Ramasco. 'Systematic comparison of trip distribution laws and models'. In: *Journal of Transport Geography* 51 (Feb. 2016), pp. 158–169. ISSN: 09666923. DOI: 10.1016/j.jtrangeo.2015.12.008.

[59] Maxime Lenormand et al. 'A Universal Model of Commuting Networks'. In: *PLoS ONE* 7.10 (Oct. 2012). Ed. by Renaud Lambiotte, e45985. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0045985.

[60] Chia-Chen Liu et al. 'Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy'. In: *Nature Reviews Neurology* 9.2 (Feb. 2013), pp. 106–118. ISSN: 1759-4758, 1759-4766. DOI: 10.1038/nrneurol.2012.263.

[61] Xin Liu, Tsuyoshi Murata and Ken Wakita. 'Detecting network communities beyond assortativity-related attributes'. In: *Physical Review E* 90.1 (July 2014). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.90.012806.

[62] Lorenzo Lucchini et al. 'Living in a pandemic: changes in mobility routines, social activity and adherence to COVID-19 protective measures'. In: *Scientific Reports* 11.1 (Dec. 2021). ISSN: 2045-2322. DOI: 10.1038/s41598-021-04139-1.

[63] Ulrike von Luxburg. 'A tutorial on spectral clustering'. In: *Statistics and Computing* 17.4 (Dec. 2007), pp. 395–416. ISSN: 0960-3174, 1573-1375. DOI: 10.1007/s11222-007-9033-z.

[64] Mahendra Mariadassou, Stéphane Robin and Corinne Vacher. 'Uncovering latent structure in valued graphs: A variational approach'. In: *The Annals of Applied Statistics* 4.2 (June 2010), pp. 715–742. ISSN: 1932-6157. DOI: 10.1214/10-AOAS361.

[65] A. Paolo Masucci et al. 'Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows'. In: *Physical Review E* 88.2 (Aug. 2013). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.88.022812.

[66] Mattia Mazzoli et al. 'Field theory for recurrent mobility'. In: *Nature Communications* 10.1 (Dec. 2019), p. 10. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11841-2.

[67] Marina Meilă. 'Comparing clusterings-an information based distance'. In: *Journal of Multivariate Analysis* 98.5 (2007), pp. 873–895. ISSN: 0047-259X. DOI: 10.1016/j.jmva.2006.11.013.

[68] P. J. Mucha et al. 'Community Structure in Time-Dependent, Multiscale, and Multiplex Networks'. In: *Science* 328.5980 (May 2010), pp. 876–878. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1184819.

[69] Dakota Murray et al. 'Unsupervised embedding of trajectories captures the latent structure of mobility'. In: *arXiv:2012.02785 [physics]* (Dec. 2020). arXiv: 2012.02785.

[70] Zachary Neal. 'Identifying statistically significant edges in one-mode projections'. In: *Social Network Analysis and Mining* 3.4 (Dec. 2013), pp. 915–924. ISSN: 1869-5450, 1869-5469. DOI: 10.1007/s13278-013-0107-y.

[71] M. E. J. Newman. 'Equivalence between modularity optimization and maximum likelihood methods for community detection'. In: *Physical Review E* 94.5 (Nov. 2016). ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.94.052315.

[72] M. E. J. Newman. 'Fast algorithm for detecting community structure in networks'. In: *Physical Review E* 69.6 (June 2004). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.69.066133.

[73] M. E. J. Newman. 'Finding community structure in networks using the eigenvectors of matrices'. In: *Physical Review E* 74.3 (Sept. 2006). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.74.036104.

[74] M. E. J. Newman, George T. Cantwell and Jean-Gabriel Young. 'Improved mutual information measure for clustering, classification, and community detection'. In: *Physical Review E* 101.4 (Apr. 2020), p. 042304. ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.101.042304.

[75] M. E. J. Newman and M. Girvan. 'Finding and evaluating community structure in networks'. In: *Physical Review E* 69.2 (Feb. 2004), p. 026113. DOI: 10.1103/PhysRevE.69.026113.

[76] M. E. J. Newman and Gesine Reinert. 'Estimating the Number of Communities in a Network'. In: *Physical Review Letters* 117.7 (Aug. 2016). ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.117.078301.

[77] Mark Newman. *Networks*. Oxford University Press, Oct. 2018. ISBN: 978-0-19-880509-0. DOI: 10.1093/oso/9780198805090.001.0001.

[78] Krzysztof Nowicki and Tom A. B Snijders. 'Estimation and Prediction for Stochastic Blockstructures'. In: *Journal of the American Statistical Association* 96.455 (Sept. 2001), pp. 1077–1087. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/016214501753208735.

[79] Nuria Oliver et al. 'Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle'. In: *Science Advances* 6.23 (June 2020). ISSN: 2375-2548. DOI: 10.1126/sciadv.abc0764.

[80] World Health Organization. *Dementia — Key facts*. Sept. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/dementia (visited on 18/01/2022).

[81] A. Roxana Pamfil et al. 'Relating Modularity Maximization and Stochastic Block Models in Multilayer Networks'. In: *SIAM Journal on Mathematics of Data Science* 1.4 (Jan. 2019), pp. 667–698. ISSN: 2577-0187. DOI: 10.1137/18M1231304.

[82] Adina Roxana Pamfil. 'Communities in annotated, multilayer, and correlated networks'. PhD Thesis. University of Oxford, 2018.

[83] Lia Papadopoulos et al. 'Comparing two classes of biological distribution systems using network analysis'. In: *PLOS Computational Biology* 14.9 (Sept. 2018). Ed. by Samuel J. Gershman, e1006428. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006428.

[84] Leto Peel, Daniel B. Larremore and Aaron Clauset. 'The ground truth about metadata and community detection in networks'. In: *Science Advances* 3.5 (May 2017), e1602548. ISSN: 2375-2548. DOI: 10.1126/sciadv.1602548.

[85] Tiago P. Peixoto. 'Bayesian Stochastic Blockmodelling'. In: *Advances in Network Clustering and Blockmodeling*. Ed. by Patrick Doreian, Vladimir Batagelj and Anuška Ferligoj. Wiley, 2020, pp. 289–332. ISBN: 978-1-119-48329-8 978-1-119-22468-6 978-1-119-22467-9.

[86] Tiago P. Peixoto. 'Disentangling Homophily, Community Structure, and Triadic Closure in Networks'. In: *Physical Review X* 12.1 (Jan. 2022), p. 011004. ISSN: 2160-3308. DOI: 10.1103/PhysRevX.12.011004.

[87] Tiago P. Peixoto. 'Inferring the mesoscale structure of layered, edge-valued, and time-varying networks'. In: *Physical Review E* 92.4 (Oct. 2015), p. 042807. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.92.042807.

[88] Tiago P. Peixoto. 'Nonparametric Bayesian inference of the microcanonical stochastic block model'. In: *Physical Review E* 95.1 (Jan. 2017). ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.95.012317.

[89] Tiago P. Peixoto. 'Nonparametric weighted stochastic block models'. In: *Physical Review E* 97.1 (Jan. 2018), p. 012306. ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.97.012306.

[90] Tiago P. Peixoto. 'Parsimonious module inference in large networks'. In: *Physical Review Letters* 110.14 (Apr. 2013). ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.110.148701.

[91] Tiago P. Peixoto. 'The graph-tool python library'. In: *figshare* (2014). DOI: 10.6084/m9.figshare.1164194.

[92] Diego Pennacchioli et al. 'Explaining the product range effect in purchase data'. In: *2013 IEEE International Conference on Big Data*. IEEE, Oct. 2013, pp. 648–656. ISBN: 978-1-4799-1293-3. DOI: 10.1109/BigData.2013.6691634.

[93] Diego Pennacchioli et al. 'The retail market as a complex system'. In: *EPJ Data Science* 3.1 (Dec. 2014). ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-014-0033-x.

[94] Duccio Piovani et al. 'Measuring accessibility using gravity and radiation models'. In: *Royal Society Open Science* 5.9 (Sept. 2018), p. 171668. ISSN: 2054-5703, 2054-5703. DOI: 10.1098/rsos.171668.

[95] Mason A. Porter, Jukka-Pekka Onnela and Peter J. Mucha. 'Communities in Networks'. In: *American Mathematical Society* 56.9 (2009), pp. 1082–1097. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2009.11.002.

[96] Vyas Ramasubramani et al. 'signac: A Python framework for data and workflow management'. In: *Proceedings of the 17th Python in Science Conference*. 2018, pp. 152–159. DOI: 10.25080/Majora-4af1f417-016.

[97] E. G. Ravenstein. 'The Laws of Migration'. In: *Journal of the Statistical Society of London* 48.2 (June 1885), p. 167. ISSN: 09595341. DOI: 10.2307/2979181.

[98] Jörg Reichardt and Stefan Bornholdt. 'Statistical mechanics of community detection'. In: *Physical Review E* 74.1 (July 2006). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.74.016110.

[99] William J. Reilly. *The law of retail gravitation*. New York: W.J. Reilly, 1931.

[100] Maria A. Riolo et al. 'Efficient method for estimating the number of communities in a network'. In: *Physical Review E* 96.3 (Sept. 2017). ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.96.032310.

[101] M. Puck Rombach et al. 'Core-Periphery Structure in Networks'. In: *SIAM Journal on Applied Mathematics* 74.1 (Jan. 2014), pp. 167–190. ISSN: 0036-1399, 1095-712X. DOI: 10.1137/120881683.

[102] M. Rosvall, D. Axelsson and C. T. Bergstrom. 'The map equation'. In: *The European Physical Journal Special Topics* 178.1 (Nov. 2009), pp. 13–23. ISSN: 1951-6355, 1951-6401. DOI: 10.1140/epjst/e2010-01179-1.

[103] Martin Rosvall et al. 'Different Approaches to Community Detection'. In: *Advances in Network Clustering and Blockmodeling*. Ed. by Patrick Doreian, Vladimir Batagelj and Anuška Ferligoj. Wiley, 2020, pp. 105–119. ISBN: 978-1-119-22470-9 978-1-119-48329-8.

[104] Wang Ru and Chi Li-Ping. 'Opinion Dynamics on Complex Networks with Communities'. In: *Chinese Physics Letters* 25.4 (Apr. 2008), pp. 1502–1505. ISSN: 0256-307X, 1741-3540. DOI: 10.1088/0256-307X/25/4/091.

[105] Clodomir Santana et al. 'Changes in the time-space dimension of human mobility during the COVID-19 pandemic'. In: (Jan. 2022). DOI: 10.48550/arXiv.2201.06527.

[106] Marta Sarzynska et al. 'Null models for community detection in spatially embedded, temporal networks'. In: *Journal of Complex Networks* 4.3 (Sept. 2016), pp. 363–406. ISSN: 2051-1310, 2051-1329. DOI: 10.1093/comnet/cnv027.

[107] Michael T. Schaub and Leto Peel. 'Hierarchical community structure in networks'. In: (Sept. 2020). arXiv:2009.07196 [cs].

[108] Michael T. Schaub, Santiago Segarra and John N. Tsitsiklis. 'Blind Identification of Stochastic Block Models from Dynamical Observations'. In: *SIAM Journal on Mathematics of Data Science* 2.2 (Jan. 2020), pp. 335–367. ISSN: 2577-0187. DOI: 10.1137/19M1263340.

[109] Frank Schlosser et al. 'COVID-19 lockdown induces disease-mitigating structural changes in mobility networks'. In: *Proceedings of the National Academy of Sciences* 117.52 (Dec. 2020), pp. 32883–32890. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2012326117.

[110] M. A. Serrano, M. Boguñá and A. Vespignani. 'Extracting the multiscale backbone of complex weighted networks'. In: *Proceedings of the National Academy of Sciences* 106.16 (Apr. 2009), pp. 6483–6488. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0808904106.

[111] Filippo Simini et al. 'A universal model for mobility and migration patterns'. In: *Nature* 484.7392 (Feb. 2012), pp. 96–100. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature10856.

[112] Richard Sinkhorn. 'A Relationship Between Arbitrary Positive Matrices and Doubly Stochastic Matrices'. In: *Annals of Mathematical Statistics* 35.2 (June 1964). Publisher: Institute of Mathematical Statistics, pp. 876–879. ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177703591.

[113] Richard Sinkhorn and Paul Knopp. 'Concerning nonnegative matrices and doubly stochastic matrices'. In: *Pacific Journal of Mathematics* 21.2 (May 1967), pp. 343–348. ISSN: 0030-8730, 0030-8730. DOI: 10.2140/pjm.1967.21.343.

[114] Il'ya Meerovich Sobol'. 'On the distribution of points in a cube and the approximate evaluation of integrals'. In: *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* 7.4 (1967). Publisher: Russian Academy of Sciences, Branch of Mathematical Sciences, pp. 784–802.

[115] Alzheimer's Society. *Alzheimer's Society's view on demography*. June 2020. URL: https://www.alzheimers.org.uk/about-us/policy-and-influencing/what-we-think/demography (visited on 18/01/2022).

[116] Chaoming Song et al. 'Modelling the scaling properties of human mobility'. In: *Nature Physics* 6.10 (Oct. 2010), pp. 818–823. ISSN: 1745-2473, 1745-2481. DOI: 10.1038/nphys1760.

[117] Samuel A. Stouffer. 'Intervening Opportunities: A Theory Relating Mobility and Distance'. In: *American Sociological Review* 5.6 (1940), pp. 845–867. ISSN: 0003-1224. DOI: 10.2307/2084520.

[118] Atsushi Tero et al. 'Rules for Biologically Inspired Adaptive Network Design'. In: *Science* 327.5964 (Jan. 2010), pp. 439–442. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1177894.

[119] Amanda L. Traud et al. 'Comparing Community Structure to Characteristics in Online Collegiate Social Networks'. In: *SIAM Review* 53.3 (Jan. 2011), pp. 526–543. ISSN: 0036-1445, 1095-7200. DOI: 10.1137/080734315.

[120] Alzheimer's Research UK. *Sea Hero Quest*. Sept. 2021. URL: https://www.alzheimersresearchuk.org/research/for-researchers/resources-and-information/sea-hero-quest/ (visited on 21/01/2022).

[121] Toni Vallès-Català et al. 'Multilayer Stochastic Block Models Reveal the Multilayer Structure of Complex Networks'. In: *Physical Review X* 6.1 (Mar. 2016), p. 011036. ISSN: 2160-3308. DOI: 10.1103/PhysRevX.6.011036.

[122] Pauli Virtanen et al. 'SciPy 1.0: fundamental algorithms for scientific computing in Python'. In: *Nature Methods* 17.3 (Mar. 2020), pp. 261–272. ISSN: 1548-7091, 1548-7105. DOI: 10.1038/s41592-019-0686-2.

[123] Robert S. Weiss and Eugene Jacobson. 'A Method for the Analysis of the Structure of Complex Organizations'. In: *American Sociological Review* 20.6 (Dec. 1955), p. 661. ISSN: 00031224. DOI: 10.2307/2088670.

[124] A G Wilson. 'A Family of Spatial Interaction Models, and Associated Developments'. In: *Environment and Planning A: Economy and Space* 3.1 (Mar. 1971), pp. 1–32. ISSN: 0308-518X, 1472-3409. DOI: 10.1068/a030001.

[125] A G Wilson. 'Comments on Alonso's 'Theory of Movement''. In: *Environment and Planning A: Economy and Space* 12.6 (June 1980), pp. 727–732. ISSN: 0308-518X, 1472-3409. DOI: 10.1068/a120727.

[126] Yingxiang Yang et al. 'Limits of Predictability in Commuting Flows in the Absence of Data for Calibration'. In: *Scientific Reports* 4.1 (2014). ISSN: 2045-2322. DOI: 10.1038/srep05662.

[127] Fabian Ying et al. 'Customer mobility and congestion in supermarkets'. In: *Physical Review E* 100.6 (Dec. 2019), p. 062304. ISSN: 2470-0045, 2470-0053. DOI: 10.1103/PhysRevE.100.062304.

[128] Wayne W. Zachary. 'An Information Flow Model for Conflict and Fission in Small Groups'. In: *Journal of Anthropological Research* 33.4 (Dec. 1977), pp. 452–473. ISSN: 0091-7710. DOI: 10.1086/jar.33.4.3629752.

[129] Tao Zhou et al. 'Bipartite network projection and personal recommendation'. In: *Physical Review E* 76.4 (Oct. 2007). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.76.046115.

[130] George Kingsley Zipf. 'The P1 P2/D Hypothesis: On the Intercity Movement of Persons'. In: *American Sociological Review* 11.6 (Dec. 1946), pp. 677–686.

Appendix A

Recovery using Spatial modularity

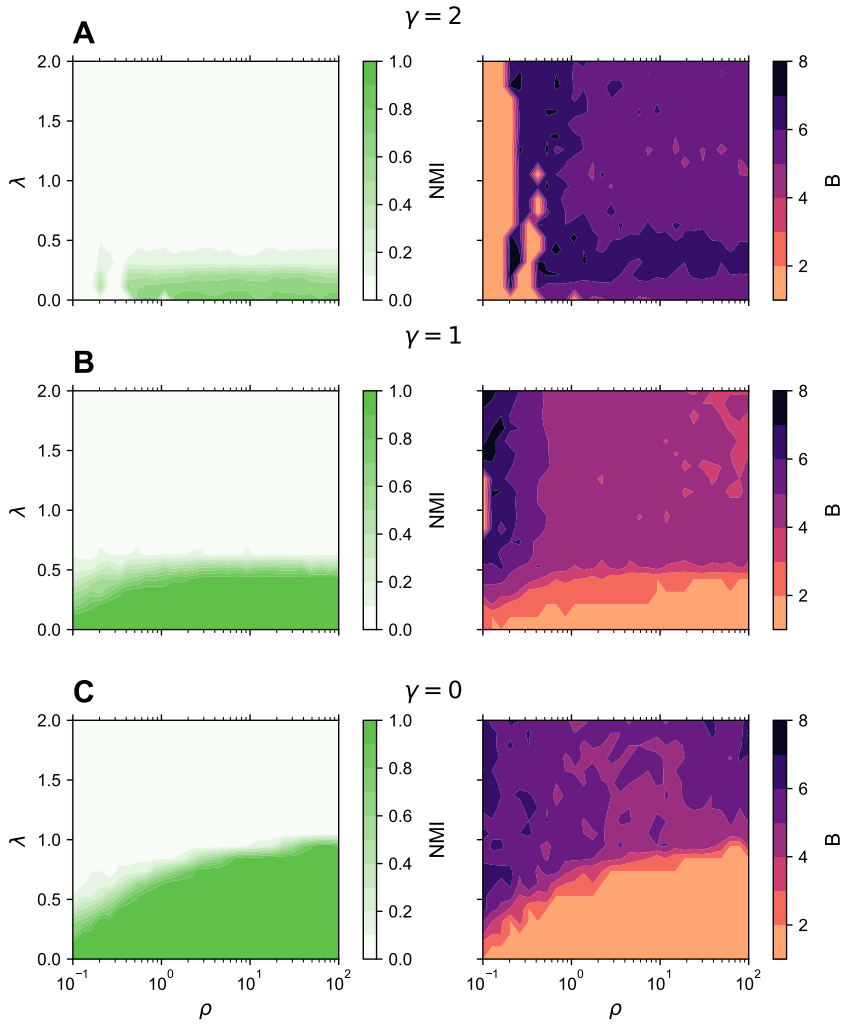


Figure A.1: Phase space of the Sobol layout for the Spatial modularity introduced by Expert et al. [27], as measured with the average NMI of the recovered partition (left) and the average number of blocks (right). For each (ρ, λ) pair, we generated 10 different networks (the output of the modularity optimisation routine is deterministic).

Appendix B

Uniform and Exponential models

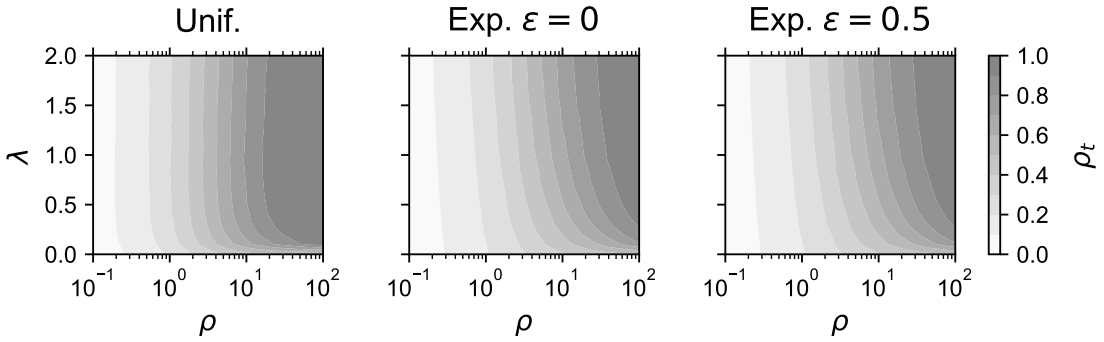


Figure B.1: Topological density for the Uniform and Exponential layouts (both fully correlated and uncorrelated); the networks are created with $N = 80$ nodes.

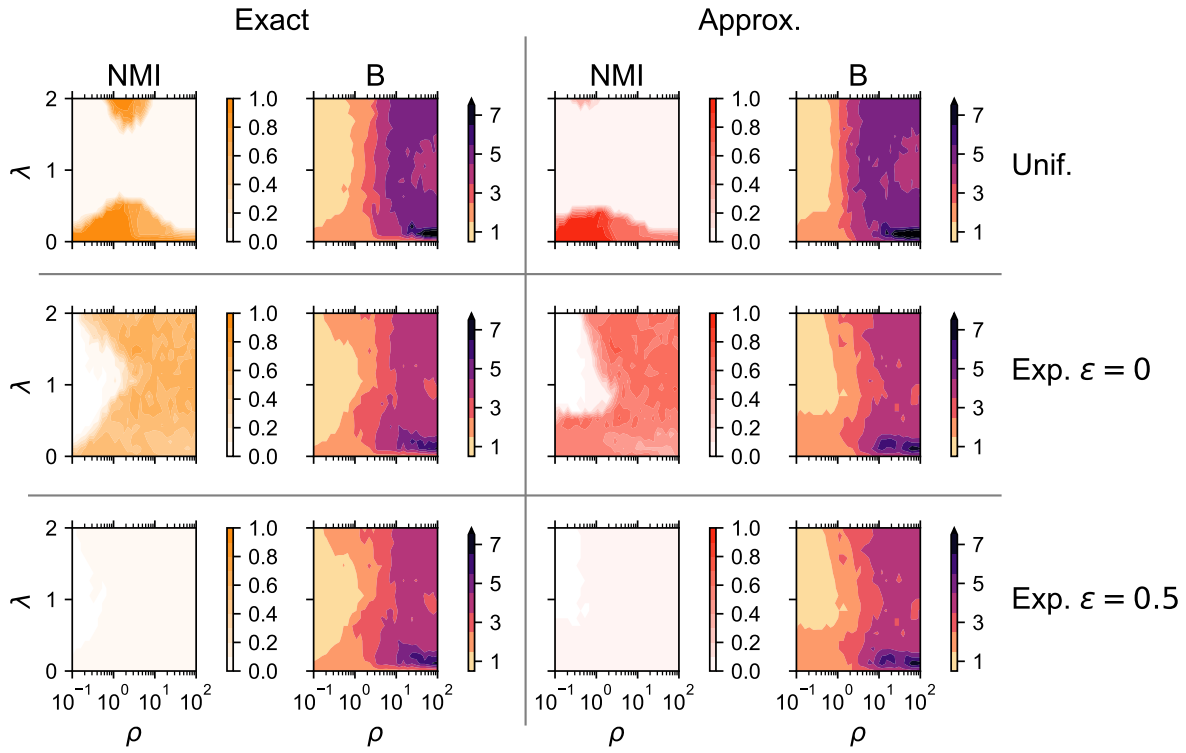


Figure B.2: Uniform and Exponential layouts (both fully correlated and uncorrelated). The recovery employs the radiation model.

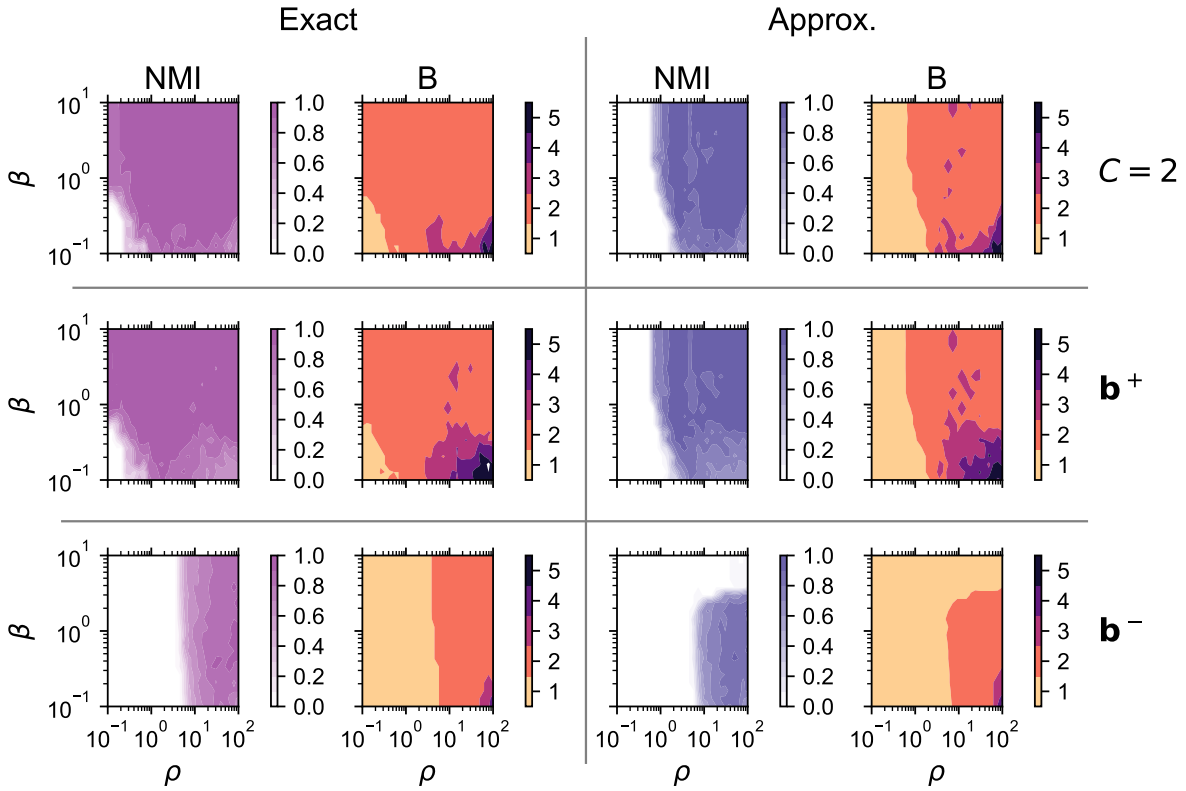


Figure B.3: Correlated Group Membership (exponential) layout with exponential weights. The block assignment is fully correlated with space ($\epsilon = 0$). The best results are obtained for the gravity model (pictured) and the layered SBM (top row).

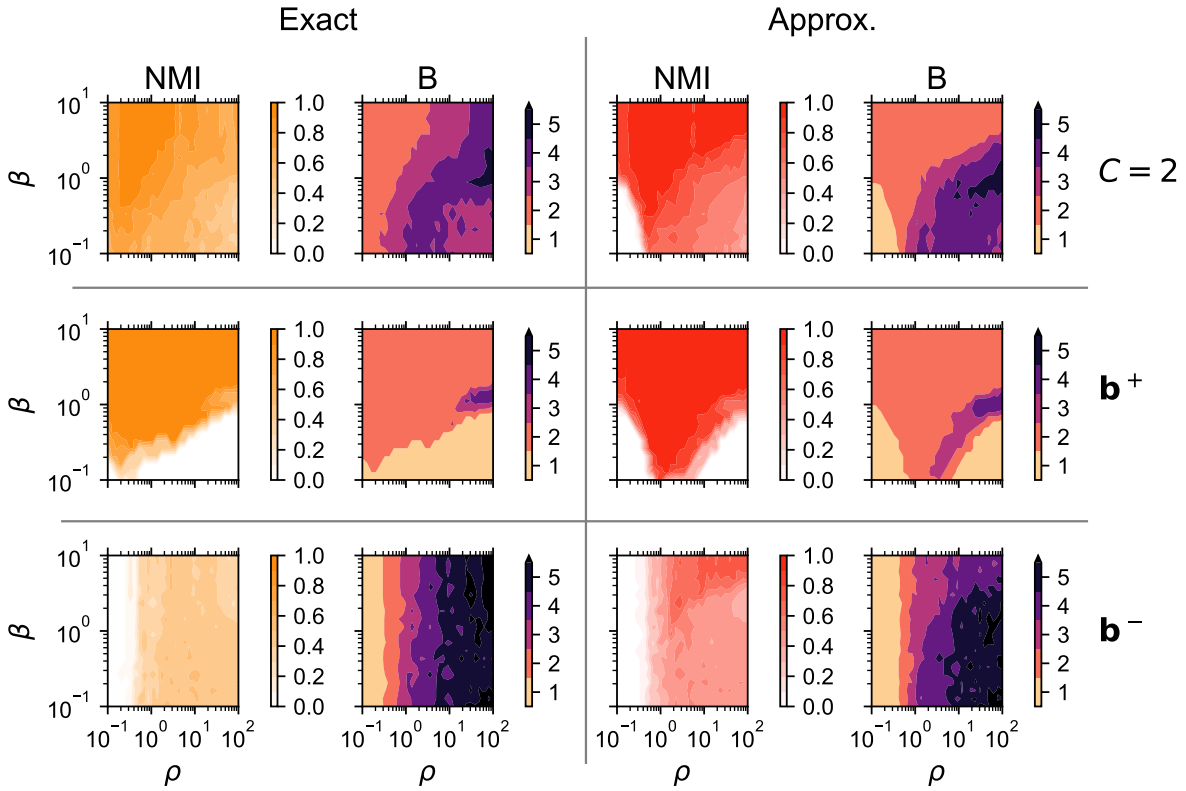


Figure B.4: Correlated Group Membership (exponential) layout with exponential weights and recovery based on the radiation model. The block assignment is fully correlated with space ($\epsilon = 0$).

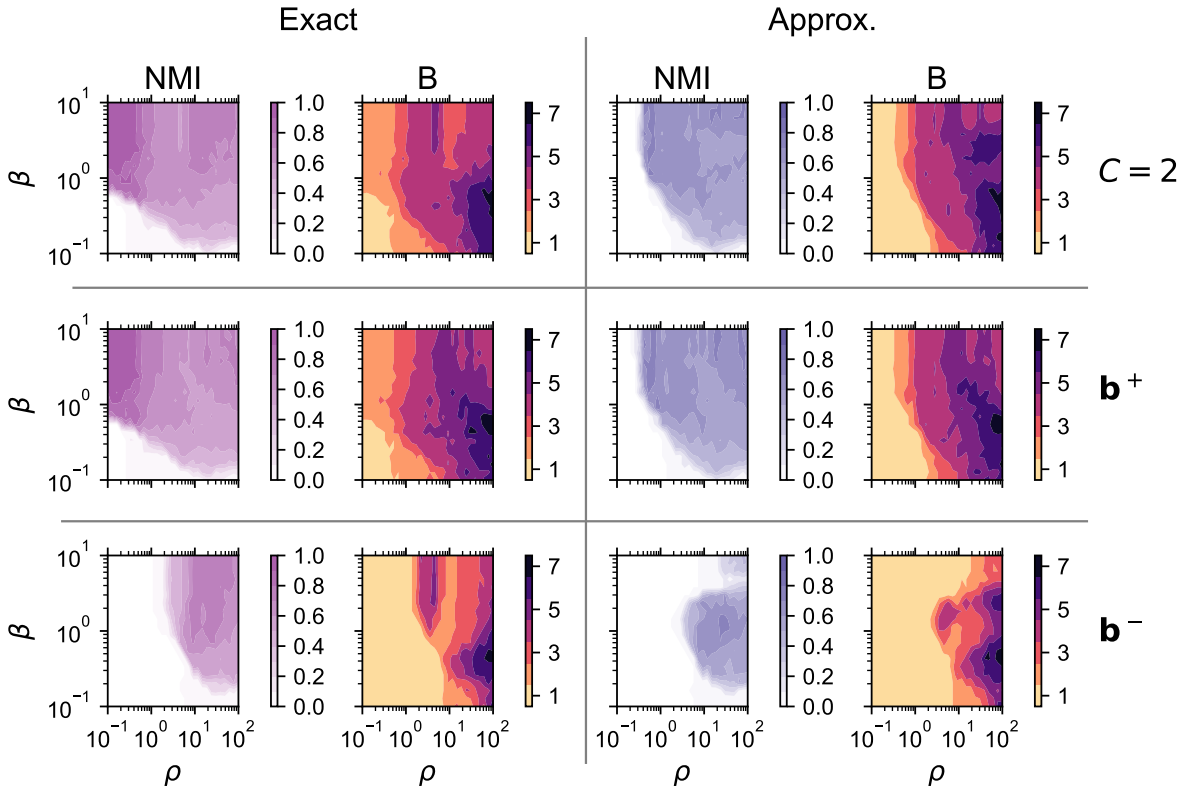
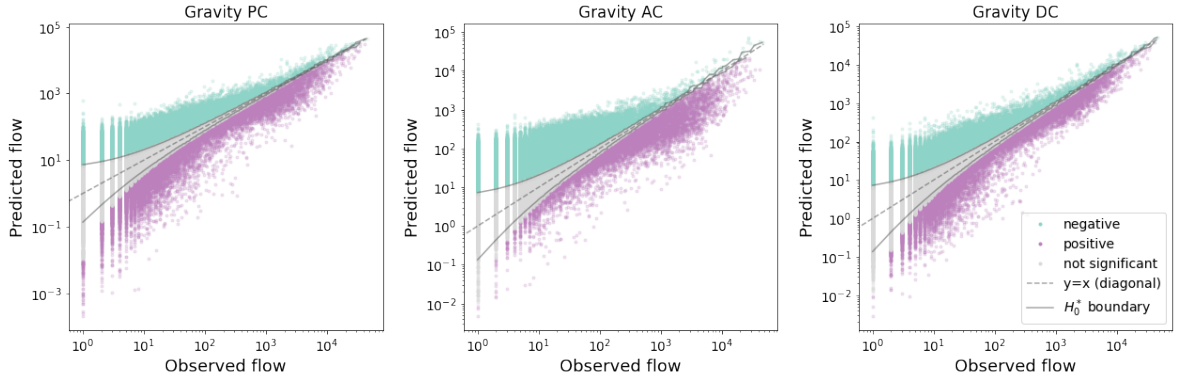


Figure B.5: Correlated Group Membership (exponential) layout with exponential weights and recovery based on the gravity model. The block assignment is not correlated with with space ($\epsilon = 0.5$).

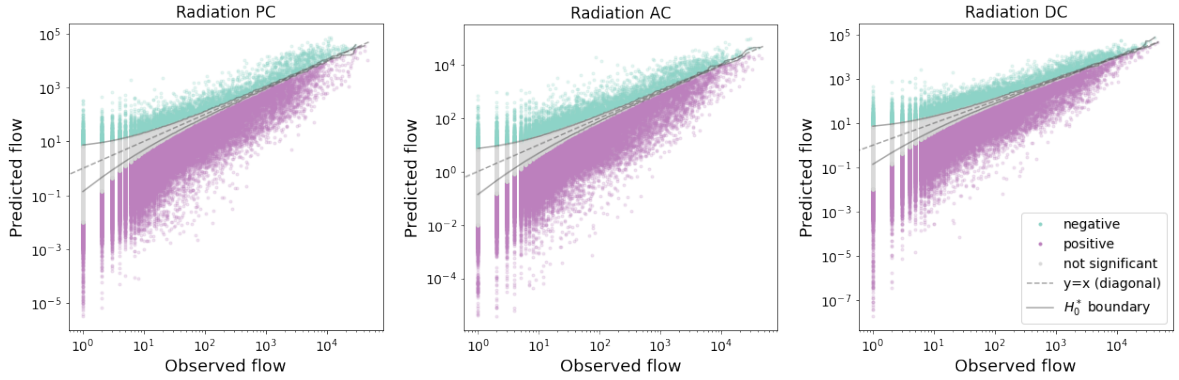
Appendix C

Number of edges in the commuting spatial backbones



(a) (b) (c)

Figure C.1: Gravity model: (a) Production-constrained (b) attraction-constrained and (c) doubly-constrained



(a) (b) (c)

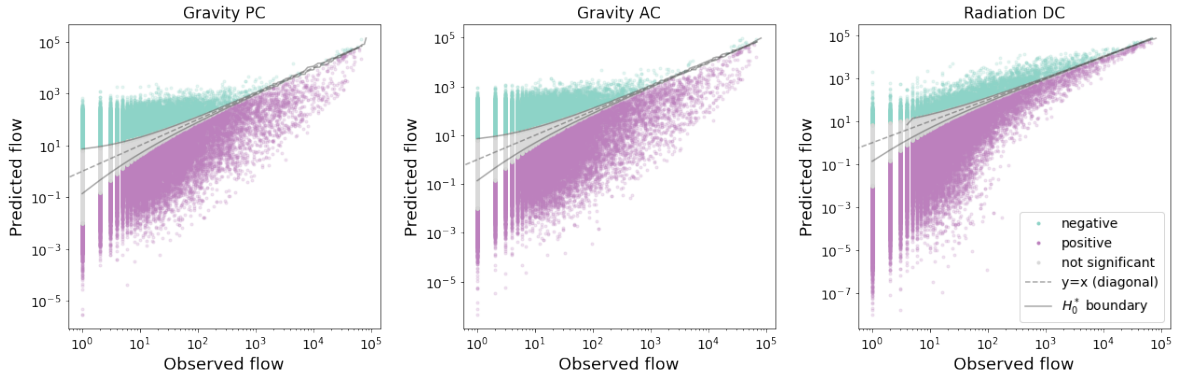
Figure C.2: Radiation model: (a) Production-constrained (b) attraction-constrained and (c) doubly-constrained

Model		Test	Positive	Negative	Not-significant	Total
Grav.	PC	exact	11,533	43,449	53,160	108,142
		approx.	14,758	42,035	51,349	
	AC	exact	6,947	73,342	27,853	
		approx.	7,579	71,947	28,616	
	DC	exact	10,309	37,018	60,815	
		approx.	13,308	35,606	59,228	
Rad.	PC	exact	56,037	4,931	47,174	108,142
		approx.	74,123	4,811	29,208	
	AC	exact	53,390	5,923	48,829	
		approx.	71,989	5,783	30,370	
	DC	exact	52,020	6,233	49,890	
		approx.	69,504	6,045	32,593	

Table C.1: Number of significant edges (at a 0.01 significance level) for the different models applied to the network of commuting.

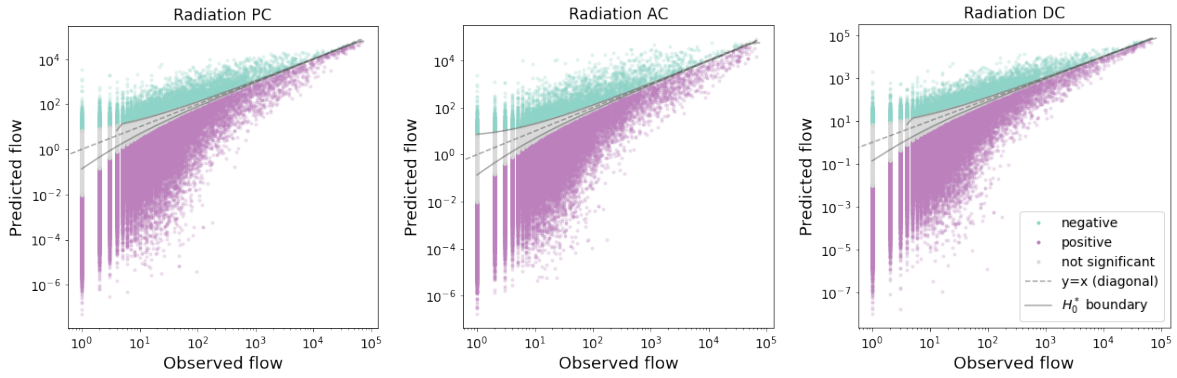
Appendix D

Number of edges in the retail spatial backbones



(a) (b) (c)

Figure D.1: Gravity model: (a) Production-constrained (b) attraction-constrained and (c) doubly-constrained



(a) (b) (c)

Figure D.2: Radiation model: (a) Production-constrained (b) attraction-constrained and (c) doubly-constrained

Model		Test	Positive	Negative	Not-significant	Total
Grav.	PC	exact	262,515	544,500	825,744	1,632,759
		approx.	434,339	526,799	671,621	
	AC	exact	316,001	513,844	802,914	
		approx.	500,293	497,259	635,207	
	DC	exact	545,669	164,689	922,401	
		approx.	783,704	156,336	692,699	
Rad.	PC	exact	878,367	63,744	690,648	1,632,759
		approx.	1,237,375	61,898	333,486	
	AC	exact	891,091	54,965	686,703	
		approx.	1,273,919	53,396	305,444	
	DC	exact	866,276	62,491	703,992	
		approx.	1,218,880	60,547	353,332	

Table D.1: Number of significant edges (at a 0.01 significance level) for the different models applied to the retail network.

Appendix E

Median scores by age for the new navigation metrics

Figure E.1: Median and inter-quartile spread by age for level 8 (the two genders are grouped together).

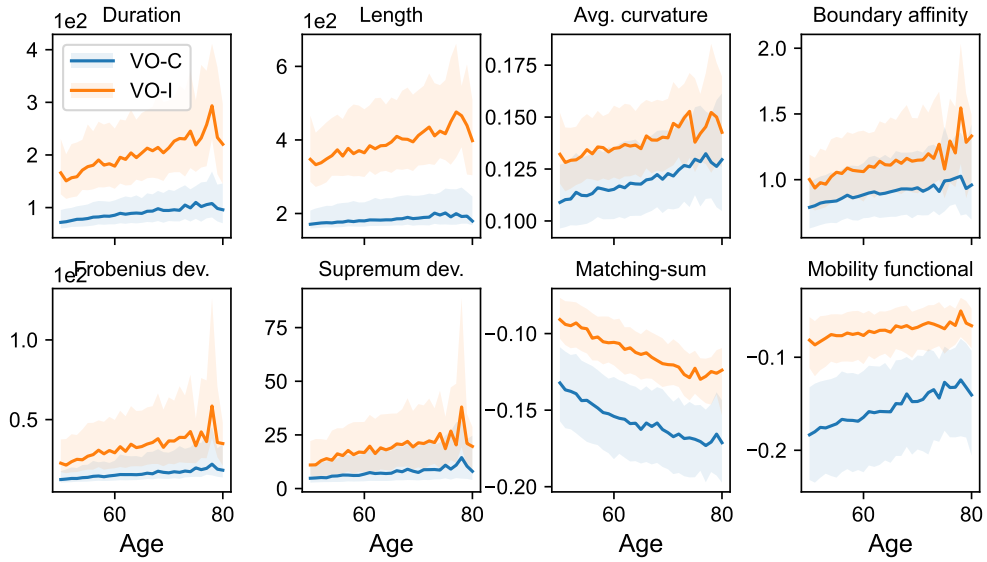


Figure E.2: Median and inter-quartile spread by age for level 11.