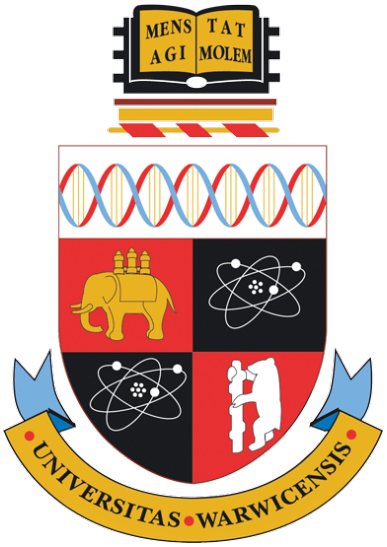
WARWICK
THE UNIVERSITY OF WARWICK



A Bayesian spatial interaction framework for optimal facility location in urban environments

by

Shanaka Perera

Thesis

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy

Department of Computer Science

January 2022

Contents

List of Tables

List of Figures

Acknowledgments

I am grateful to my supervisor, Professor Theo Damoulas, for his guidance, support and encouragement to constantly thrive for the best. I have greatly benefited from his technical and research expertise to develop as a researcher. Huge thanks to him for accepting me as his student at an important stage of my journey and helping me to get through the challenging times.

Special thanks to Professor Stephen Jarvis for making my path to engage in research at the computer science department and introducing some great opportunities. I am thankful to Professor Joao Porto de Albuquerque for his advice given during the research.

Big thanks to my colleague Virginia for collaborating with me and sharing her knowledge and technical skills. A huge thanks to Paul Davis, the director at Nimbus property technology, for allowing me to pursue an interesting research topic and valuable insights into the property industry and providing access to a comprehensive property database.

All the support from my family has been amazing to get through very challenging times over the past few years. Finally, I would like to thank my friends from the PhD cohort and Henry, Nuzhi, Shanika, Arun and all who supported in making it an interesting journey.

Declarations

1 Publications

I hereby declare that this dissertation is original work and has not been submitted for a degree or diploma or other qualification at any other university. Parts of this thesis have been previously published or currently under review:

[116] Shanaka Perera, Theo Damoulas, Paul Davis, and Stephen Jarvis. Modelling business rates in england with big spatial data. In *In Proceedings of SIGKDD '19: International Workshop on Urban Computing*. SIGKDD, 2019

Under Review

[117] Shanaka Perera, Virginia Aglietti, and Theodoros Damoulas. On the competitive facility location problem with an extended Bayesian spatial interaction model. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, Submitted on Dec 2021

[118] Shanaka Perera, Virginia Aglietti, and Theodoros Damoulas. A Bayesian spatial interaction model for estimating revenue and demand at business facilities. *Nature Computational Science*, Submitted on Jan 2022

Research was performed in collaboration during the development of this thesis, but does not form part of the thesis:

[154] Godwin Yeboah, João Porto de Albuquerque, Rafael Troilo, Grant Tregonning, Shanaka Perera, Syed AK Ahmed, Motunrayo Ajisola, Ornob Alam, Navneet Aujla, Syed Iqbal Azam, et al. Analysis of openstreetmap data quality at different stages of a participatory mapping process: Evidence from slums in africa and asia. *ISPRS International Journal of Geo-Information*, 10(4):265, 2021

2 Sponsorships and Grants

This research was funded by:

- Sponsor 1 : Engineering and Physical Sciences Research Council (EPSRC) National Productivity Investment Fund (NPIF) (EP/R512229/1). 2017-2022.

- Sponsor 2 : Assured Property Group, Innovation Center, Warwick Innovation Center, Gallows Hill, Warwick, CV34 6UW. 2017-2020.

Abstract

The actions of many interacting entities within socio-economic systems proclaim the configurations such as the spatial structure in urban environments. Hence understanding these underlying interactions are important in making the location decisions for growth in urban systems. In this thesis, a Bayesian spatial interaction model, henceforth BSIM, is developed to provide probabilistic predictions about revenues generated by a particular business location, based on its features and the potential customers' characteristics in a given region. BSIM explicitly accounts for the competition among the facilities through a probability determined by evaluating a store-specific Gaussian distribution at a particular customer location. I propose a scalable variational inference framework that exhibits comparable performance in terms of parameter identification and uncertainty quantification while being significantly faster than competing Markov Chain Monte Carlo inference schemes.

The advantages of the introduced BSIM are explored in addressing the competitive facility location problem that typically arises when businesses plan to enter a new market or expand their presence in an environment with existing competitors. A mathematical modelling framework is formulated to simultaneously identify the location and design of new stores in order to maximise the revenue predicted from BSIM in a geographical region. Solving the underlying optimisation problem requires the provision of an exhaustive set of potential sites, which is difficult in practice. Instead, a search algorithm is proposed based on the quadtree method to overcome this challenge by hierarchically exploring geographic regions of varying spatial resolution.

This thesis introduces multiple large-scale real-world datasets compiled with open and proprietary data. Finally, demonstrate the proposed framework

by producing optimal facility locations and corresponding designs for two case studies in the supermarket and pub sectors in Greater London, providing valuable insights for planning and decision-making under uncertainty.

Acronyms

R^2 R-squared.

AUC Area Under the Curve.

BSIM Bayesian Spatial Interaction Model.

CI Credible Interval.

CV Cross-Validation.

ELBO Evidence Lower Bound.

ell expected log likelihood.

FRK Fixed Rank Kriging.

IPPP Inhomogeneous Poisson Points Process.

KL Kullback-Leiber.

LSOA Lower Layer Super Output Areas.

MCKP Multiple-Choice Knapsack Problem.

MCMC Markov Chain Monte Carlo.

MH Metropolis-Hastings.

MSE Mean-Squared Error.

NRMSE Normalised Root-Mean-Squared Error.

PDF Probability Density Function.

RMSE Root-Mean-Squared Error.

SD Standard Deviation.

SKCV Spatial k-fold cross-validation.

UK United Kingdom.

VI Variational Inference.

VOA Valuation Office Agency.

Chapter 1

Introduction

Understanding the mechanics of complex systems such as cities is a challenging task in contemporary science [6]. Cities are shaped by the actions of many interacting individuals within socio-economic systems. These interactions decide the existence, growth and decline in urban environments [61]. In the last two decades, advancements in technology have significantly influenced these interactions and urban structures. For instance, customers' interactions with businesses have predominantly moved from physical stores to online, challenging the existence of physical retail stores. In Great Britain, online sales as a proportion of total retail sales have tripled in a decade, reaching 21% in 2019 [110]. Technology has also enabled access to much richer large-scale datasets and increased computer processing power has made it possible to calibrate complex mathematical models. Hence much attention has been gained recently to adopt the Bayesian approach to quantify aleatoric and epistemic uncertainty [84]. Indeed these developments demand a Bayesian approach for formulating interactions in urban environments for making location decisions and is the subject of this thesis.

System theories help better understand the complexities of the world we live in and describe the underlying processes of real-world problems. These theories affirm that events do not occur in isolation but involve interrelated entities that form a consolidated group known as the system [138]. Thus primarily, it aims to identify relevant interrelationships from the perspective of entities and processes [33]. The nature of these relationships determines how the entities interact with each other. The associations could be either between the same class of entities or across multiple classes and at different levels. System analysis has the great advantage of identifying underlying processes to predict possible future outcomes.

Defining systems are never the same and depend on the problem as well as the motivations of the analysis. The elements of a system may not only consist of built components (e.g. computer chips, retail stores etc.) and natural

entities (e.g. trees, rivers etc.) but also humans considered as intelligent agents, thus known as socio-economic systems [115]. In order to explain human behaviour and economics, it is important to understand the eco-dynamics of the larger interconnected general system [19]. Broadly, the economic system determines how societies or governments regulate the distribution of services and goods across a geographic region, forming the financial structure of a certain community [64].

1.1 Motivation

This thesis focuses on a significant component that drives economic systems: the interactions of intelligent agents with entities providing services or products. These social and economic systems vary across geographic space and time. For instance, over time, the entities are drawn into areas in order to benefit from positive externalities arising from shared resources, leading to changes in spatial dynamics [47]. Furthermore, at a given moment, the differences in space are evident comparing rural and metropolitan areas, where more contrasted opportunities are available to citizens [138]. This study focuses on the behaviour of intelligent agents and decision makings on entities within spatial dimensions in a certain time horizon.

System dynamics effects in respect to space are broadly explained by spatial heterogeneity, and spatial dependence [4]. Spatial heterogeneity explains the variation of the distribution in events across the geographic space [45]. For example, differences in the spending capacity across regions. In contrast, spatial dependencies define the association between events at a specific location and actions in neighbouring areas. Such relationships influence the entities to benefit from positive externalities known as agglomeration economies [99]. The interactions between the same class of entities lead to the specialism that depicts localisation economies, whereas links between cross-industries produce diversified cities forming urban agglomeration economies [47]. The knowledge created by firms or institutions does not remain within only but spill over to influence the other neighbouring economies [3]. The strategic role of interactions among agents, institutions, and local economies is of primary interest in explaining urban growth dynamics.

Henceforth a major area of research in urban science is to describe the flows that arise from interactions between entities such as humans, goods and service providers, and physical infrastructure [11]. The spatial interactions are of various sorts: trade flows, commutes, ideas, capital, tourists, transportation, migrants etc. In economic systems such as market economies, the financial decisions and pricing of goods and services are controlled by the interactions between supply served by facilities and demand created at demand points, also

referred to as customers [88]. The flows (F_{ij}) between destination (facility D_j) and origin (demand point O_i) in an urban system are illustrated in Fig. 1.1.

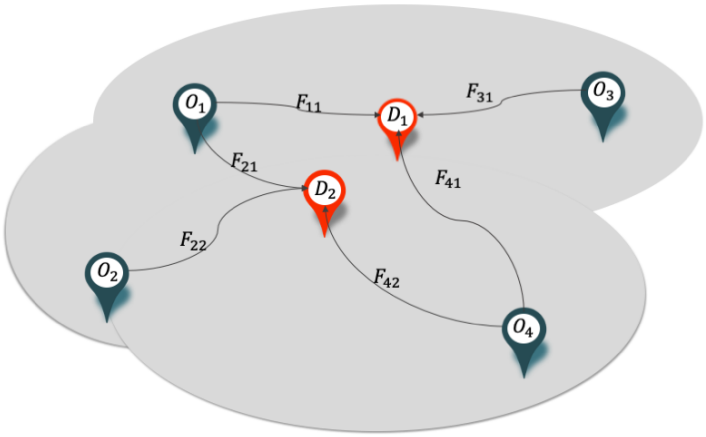


Figure 1.1: Illustration of possible flows in an urban system. It is assumed that there are 4 demand points and 2 facilities. The flow F_{ij} denotes the flow of quantities from origin O_i to destination zone D_j.

Mathematical modelling of spatial interactions has a long history dating back to 1929 [125]. Since then, many researchers have formulated the declining pattern of flows with the increase in distance similar to the gravity model [38, 79, 105]. The utility of a facility perceived by the customer depends on the location and attraction. The facilities closer to customer and with high quality are more attractive. Various methods are applied in the literature to model utility, but the most popular method is to divide the attraction by a function of distance[11]. Early studies assumed that customers prioritise the most attractive facility, known as the deterministic rule [37, 74]. However, this hypothesis did not find empirical evidence except for areas with limited shopping options, and transportation is difficult [51]. On the contrary, Huff introduced the probabilistic rule that states customers patronise all facilities, and demand is allocated proportionally to the utility [78]. Many extensions to this model are predominantly applied in forecasting future customer choices, migration patterns, and transportation demand [16, 34, 95, 135, 148]. However, accessing individual customer transactional data is highly restrictive due to the confidentiality compared to transportation or migration data between designations. Hence deriving aggregate forecasts such as revenues or demand at store levels from a disaggregate spatial interaction model remains a challenge[129].

The estimated revenue or market share at new sites is of prime importance for making location decisions in competitive environments [49]. A well-known phenomenon is that the most critical attributes of stores are location, location, and location. The competitive location problems formulate mathematical models to incorporate all factors that may affect maximising the market share

captured by new facilities [15, 40]. The objective of the optimisation model depends on the current state in the market of the company that searches for new sites. For instance, when a business with a chain of existing facilities plans to add several new stores, the objective is to increase market share captured by the chain, not just the additional site [42, 85]. The location decision provided by the facility location model is invaluable for decision-makers as locating stores require high investments and is not easily altered.

Even though there is a long history in mathematical and statistical modelling for facility location problems and spatial interaction models, uncertainty quantification has only been studied recently [46, 92]. The Bayesian framework fully quantifies the uncertainty whilst making inferences and model predictions [96]. In Bayesian settings, all the unknown quantities are modelled as random variables and uncertainty is reported by assigning probabilities to possible values and summaries with credible intervals [136]. Henceforth, this demonstrates many advantages over point estimates by addressing issues such as existence and data sensitivity. Bayesian frameworks have gained popularity with the advent of methods such as notably Markov chain Monte Carlo (MCMC) [70], and Variational Inference [81]. Advancements in computer processes and large-scale data availability have required using Bayesian methods to celebrate complex mathematical and statistical models to real-world problems. Henceforth in this thesis, a significant step is taken to model aggregate revenue or demand at facilities while accounting for the spatial interactions with customers to make location decisions under uncertainty.

1.2 Motivating Example

One of the key results from the real-world case study is presented in Fig. 1.2 to demonstrate the ability to produce rich inference from the state-of-the-art Bayesian spatial interaction model developed in this thesis. The map illustrates the clusters formed by customer zones with a common preference for a particular pub in the area. The most likely pub to visit is not necessarily within the cluster, as illustrated in Fig. 1.2(b). These inferred clusters would provide great insights for the businesses to understand their customer segments with respect to the geographic regions and their demographics. Additionally, the map demonstrates varying average probabilities of the most prefered pub across the Lower Layer Super Output Areas (LSOAs): red colour hotspots indicate high competition, for instance, central London. Furthermore, the results can be summarised at LSOA (Fig. 1.2 b) or postcode level (Fig. 1.2 c), and the preference for each pub regarding a postcode (Fig. 1.2 d). These inferences provide valuable insights into the underlying spatial interactions between facilities and customers that are vital for business planning and decision making

on their facility locations.

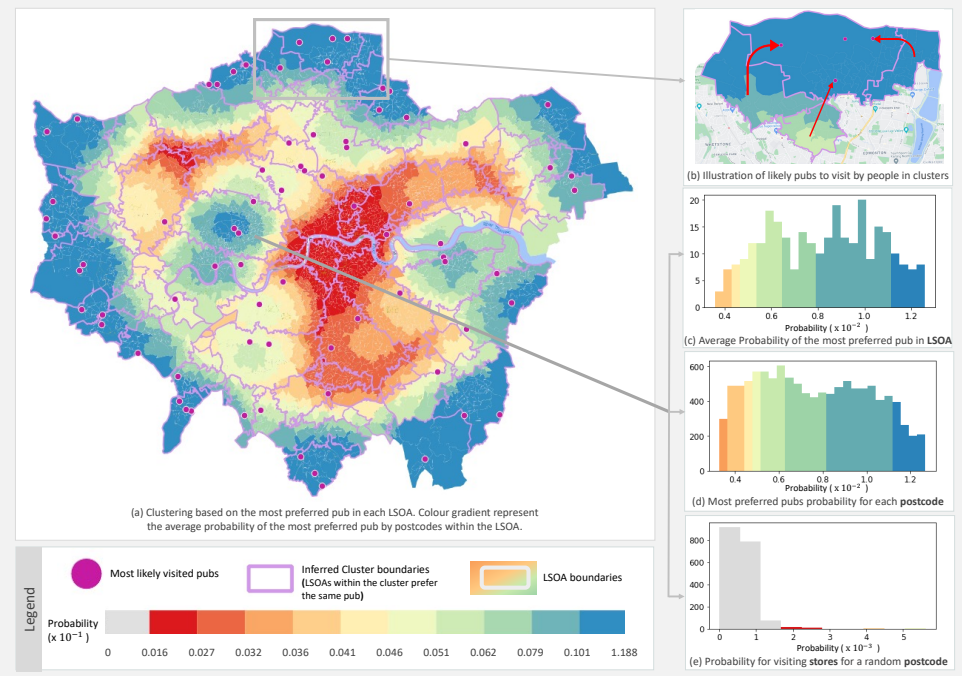


Figure 1.2: (a) Illustrates the average probabilities of the most preferred pub across the LSOAs. The most preferred pub for a postcode is the store with the highest probability of choosing. Clusters are formed from LSOAs with commonly preferred pubs. Only the clusters with ten or more LSOAs are presented. (b) Illustration of likely pubs to visit by people in clusters that are zoomed into. For the selected cluster, the histograms show: (c) the average probability of the most preferred pub in each LSOA; (d) most preferred pubs probability for each postcode; (e) probability of choosing a pub by a randomly selected postcode.

1.3 Thesis contributions

This thesis introduces mathematical models and computational techniques as well as their applications to address complex problems in urban science. A review of the literature and Bayesian methods are provided in chapter 2. The core of this thesis is the two topics forming chapters 3 and 4. This thesis reinstate the value of transforming academic research into applied outcomes for decision-makers by introducing large scale datasets and real-world case studies in chapters 5 and 6. An overview of these contributions is provided in the following subsections. Finally, conclusions and further work are discussed in chapter 7.

Chapter 3: A Bayesian spatial interaction model for estimating revenue and demand at business facilities

An important problem in socio-economic systems is modelling spatial interactions. Out of the many applications, interactions between customers and facilities have had much attention because of their role in shaping urban spatial structures. However, there is a lack of research addressing modelling with uncertainty. Additionally, a large volume of transactions is required to calibrate spatial interaction models, which is difficult to obtain. To address these shortcomings, in Chapter 3, a new spatial interaction model with a Bayesian framework is introduced. The work in this contribution makes part of the publication [118].

The key contributions are as follows:

- Developed a Bayesian spatial interaction model (BSIM) that can produce probabilistic predictions of revenues or demand generated at business facilities.
- A probabilistic method is proposed to formulates the relationship between distance and attractiveness of facilities jointly, using a facility-specific probability distribution.
- A scalable variational inference method is proposed and demonstrated its benefits compared to MCMC methods in a variety of synthetic experimental settings.

Chapter 4: On the Competitive Facility Location problem with an extended Bayesian Spatial Interaction Model

The geographical placement of a new business facility is of critical importance for commercial success. Like many spatial interaction models, BSIM also assumed fixed demand which hinders the ability to make realistic predictions. Hence BSIM is advanced and applied to a competitive facility problem to make optimal store location decisions with uncertainty. The work in this contribution makes part of the publication [117].

The key contributions are as follows:

- The BSIM is extended in order to address one of the limitations by including lost demand in competitive environments to provide more realistic revenue estimates.

- An optimisation problem is formulated to simultaneously identify optimal facility locations and corresponding designs in competitive environments and provide probability density estimates of revenues at new sites.
- A search algorithm is proposed based on the quadtree method to explore geographic regions of varying spatial resolution hierarchically. Synthetic experiments were conducted to demonstrate the performance of different sampling methods for creating potential locations.

Chapter 5: Introducing Large scale geo-spatial Datasets

The literature suffers from large scale real-world applications, mainly because acquiring granular level data is usually expensive. Hence in literature, the experiments only focus on either synthetic settings or aggregate level data. In this thesis, the most advanced datasets in the property industry are introduced with the partnership of one of the leading prop-tech companies in the UK. Three major datasets are created by combining over 20 commercial and open-source datasets. This Chapter forms parts of all three publications [116–118].

The key contributions are as follows:

- A dataset for over 1500 Pubs in Greater London is compiled to demonstrate revenue, physical store features, surrounding characteristics and customer ratings.
- A unique dataset is introduced with approximated revenues and store capacity for the nine largest supermarket chains in the UK.
- Over 150,000 postcodes, most granular administrative level, data set is compiled for Greater London to represent customer zones and characteristics.

Chapter 6: Real world applications

Demonstrating the applicability of academic research in real-world settings is challenging primarily because of the lack of scalability in models and access to relevant data. Both these shortcomings were addressed in the model building stage and by introducing granular level real-world data. To the best of my knowledge, this thesis is the first to present a fully integrated competitive facility location problem that includes both the spatial interaction modelling component and the store location optimisation framework. The work in this

contribution is taken from all three publications [116–118].

The key contributions are as follows:

- A state-of-the-art Fixed Rank Kriging model is proposed to cope with high-dimensionality and learn business rateable values from spatial context and property characteristics. By accounting for spatial effects, the model improves on current business rates valuation practice and helps with making the process more fair and transparent.
- BSIM method is applied to the pubs and supermarket sector that proved to provide the best predictive performance compared to competing approaches while providing inference at the level of customers and business facilities, delivering invaluable insights for planning and decision making.
- The optimal facility locations and their designs are demonstrated for a new company to enter the pubs' industry and expand the existence of a supermarket chain in Greater London for two industries.

1.4 Summary

This section introduced and motivated the problem that is studied in this thesis and listed the main contributions. The remaining of the thesis is organised as follows: In the next section, the essential background knowledge required to understand the remainder of this thesis is discussed. In Chapter 3, the Bayesian spatial interaction framework is introduced and is extended and develop a framework to identify the optimal location in Chapter 4. Next, in Chapter 5, large scale datasets are introduced and then apply them in real-world case studies in Chapter 6. Finally, in Chapter 7, the thesis is concluded by discussing limitations and future extensions.

Chapter 2

Background

Modelling is a mechanism that assists researchers, planners, decision-makers and many other experts in making future predictions or spatial estimations in a region. In the spatial context, its properties can be measured or reported only at a discrete set of locations on a continuous plane. However, the decision-makers are interested to know the values of uninvited places in between, hence interpolating from the known measurements. My thesis's main interest is to estimate potential returns for facilities at unexplored locations by understanding the observed measurements from the existing sites to make location decisions under uncertainty. Hence explore spatial modelling that is primarily based on distance, and spatial interaction models that explore interactions between the origin (customers) and destinations (facilities). Next, discuss the competitive facility location problem that is built on the revenue predictions. Finally, discuss the Bayesian framework and approximate inference techniques applied for model calibration in this thesis.

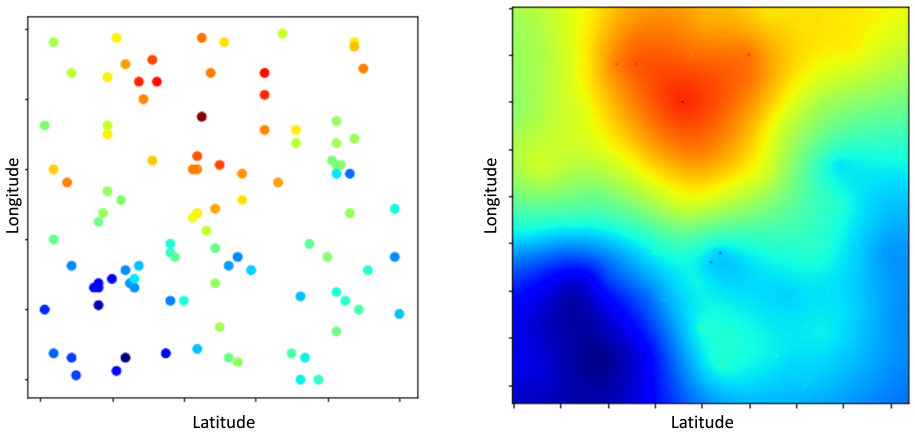
2.1 Spatial models

In spatial modelling, the focus is to construct a model based on the values measured at regular or irregular geographic locations in an area and then make estimations at any selected geo-location within the same region.

2.1.1 Kriging

Kriging is a spatial predictor in geostatistics, and synonymous with "Optimal spatial prediction" [27]. The term 'kriging' is in honour of Krige, a South African mining engineer. Kriging predicts the values of a function at a specific point by averaging the known values of the process in the neighbourhood. The spatial variability is quantified through the covariance function (or variogram). In contrast to other mathematical interpolators and regression, kriging provides estimates of the errors in its interpolations [109]. Hence can

produce maps of optimal predictions and associated prediction errors from incomplete and noisy data. Also, this method minimises the errors and predictions are unbiased, thus known as best linear unbiased prediction (BLUP). Fig. 2.1 provides an example of an synthetic spatial dataset (a) and the kriging prediction for those observations (b). The red spatial points represent high values and the blue points display low values in the simulated dataset. The same colour scheme is applied for the kriged output.



(a) (b)

Figure 2.1: A visual representation (a) of an input of spatial points and (b) Kriged output.

Kriging has become popular in the earth and environmental sciences. Drilling wells to estimate oil reserves is an expensive process, and hence obtain small samples of data and thereby interpolate [108]. However, growth in technology and the presence of remote sensing platforms on satellites have moved into massive datasets [128]. Processing large scale data with kriging is more challenging and is discussed in this section in more detail.

This section explains the Kriging formulation similar to that is demonstrated by [29]. Kriging makes inference on unobserved values of an realisation of a random process (or stochastic process or random field):

$$\{Z(s) : s \in D\}, \tag{2.1}$$

where D is a fixed subset of $\mathbb{R}^d$ with $d > 0$; meaning spatial index s varies continuously within the region D. The inferences are made using the data $Z \equiv (Z(\mathbf{s}_1) \dots Z(\mathbf{s}_n))^\top$ observed at spatial locations $\mathbf{s}_1, \dots \mathbf{s}_n$. The random process $Z(s)$ is decomposed as:

$$Z(s) = Y(s) + \varepsilon(s) \tag{2.2}$$

where $Y(\cdot) \equiv E(Z(\cdot))$ is called the large-scale variation or referred to as the

noiseless version of the Z and $\{\varepsilon(\mathbf{s})\colon \mathbf{s} \in D\}$ is a spatial white noise process distributed as $\varepsilon(\mathbf{s}) \sim N(0, \sigma^2\nu(\mathbf{s}))$ and $\nu(\mathbf{s})$ is known.

There are three popular versions of kriging: *Simple Kriging* assumes that Y is constant across the region (*first-order stationary*) and known. *Ordinary Kriging* assumes a constant unknown mean. In contrast *Universal Kriging* does not assume constant mean but is an unknown linear function. In real world settings as such in this thesis, the mean varies across the study area. Hence focus on *Universal Kriging* that assumes $Y(\cdot)$ to have a linear structure:

$$Y(\mathbf{s}) = x(s)^\top \alpha + \boldsymbol{\nu}(s), \qquad s \in D \tag{2.3}$$

where $x(\cdot)$ represents a vector process of known covariates and coefficient α are unknown. $\boldsymbol{\nu}(\cdot)$ has zero mean, $0 < \text{var}\{\boldsymbol{\nu}(s)\} < \infty, \forall s \in D$ and generally a non stationary spatial covariance function,

$$\text{cov}\{\boldsymbol{\nu}(\mathbf{u}), \boldsymbol{\nu}(\mathbf{v})\} \equiv C(\mathbf{u}, \mathbf{v}) \qquad \mathbf{u}, \mathbf{v} \in D. \tag{2.4}$$

The expression Eqs. (2.2) – (2.3) can be written as:

$$Z = \mathbf{X}\alpha + \delta, \qquad \delta = \boldsymbol{\nu} + \varepsilon \tag{2.5}$$

where $\mathbf{X}$ is a $n \times p$ matrix of known covariates $(x(\mathbf{s}_1), \ldots x(\mathbf{s}_n))^\top$ and δ is a combination of two 0 mean components, thus $E(\delta) = 0$ and $\text{var}(\delta) = \boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = \mathbf{C} + \sigma^2\mathbf{V} \tag{2.6}$$

where $C \equiv C(\mathbf{s}_i, \mathbf{s}_j)$ and $V = \text{diag}\{\nu(\mathbf{s}_1) \ldots \nu(\mathbf{s}_n)\}$. It is desired to predict Y process at location $\mathbf{s}_0, \mathbf{s}_0 \in D$:

$$\hat{Y}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)^\top \hat{\alpha} + \mathbf{k}(\mathbf{s}_0)^\top (Z - \mathbf{X}\hat{\alpha}) \tag{2.7}$$

where

$$\hat{\alpha} = (\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} Z \tag{2.8}$$

$$\mathbf{k}(\mathbf{s}_0)^\top = \mathbf{c}(\mathbf{s}_0)^\top \boldsymbol{\Sigma}^{-1} \tag{2.9}$$

and $\mathbf{c} \equiv (C(\mathbf{s}_0, \mathbf{s}_1), \ldots, C(\mathbf{s}_0, \mathbf{s}_n))^\top$. The linear unbiased predictor is obtained by minimising the root mean-squared prediction error of $\hat{Y}(\mathbf{s}_0), [E(Y(\mathbf{s}_0) - \hat{Y}(\mathbf{s}_0))]$:

$$\begin{aligned}\sigma_k(\mathbf{s}_0) = \{&C(\mathbf{s}_0, \mathbf{s}_0) - \mathbf{k}(\mathbf{s}_0)^\top \boldsymbol{\Sigma} \mathbf{k}(\mathbf{s}_0) + (x(\mathbf{s}_0) - \mathbf{X}^\top \mathbf{k}(\mathbf{s}_0))^\top \\ &(\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} (\mathbf{x}(\mathbf{s}_0) - \mathbf{X}^\top \mathbf{k}(\mathbf{s}_0))\}^{1/2}.\end{aligned} \tag{2.10}$$

For calculating both predictions (Eq. (2.7)) and prediction errors (Eq. (2.10)) requires evaluating $\mathbf{\Sigma}^{-1}$ that has a computational cost of $O(n)^3$. While powerful with small data sets, the growth in spatial Big-data poses a growing challenge to perform computations in an appropriate amount of time.

2.1.2 Fixed Rank Kriging

Chrissie and Johannesson [29] introduced the FRK model to analyse very large data sets, reducing computational cost to $O(n)$ from $O(n)^3$. In general, the covariance function is modelled as being stationary, in which case it must be a non-negative-definite function of $\mathbf{u} - \mathbf{v}$. In the FRK models the spatial dependence is captured through a set of basis functions,

$$\mathbf{S}(\mathbf{u}) \equiv (\mathbf{s}_1(\mathbf{u}), \ldots, \mathbf{s}_r(\mathbf{u}))', \qquad \mathbf{u} \in \mathbb{R}^d \tag{2.11}$$

and $\text{cov}\{\nu(\mathbf{u}), \nu(\mathbf{v})\}$ is modelled as,

$$C(\mathbf{u}, \mathbf{v}) = \mathbf{S}(\mathbf{u})'\mathbf{K}\mathbf{S}(\mathbf{v}), \qquad \mathbf{u}, \text{v} \in \mathbb{R}^d \tag{2.12}$$

where $\mathbf{K}$ is an unknown $r \times r$ symmetric positive-definite matrix. The expression Eq. (2.12) is a consequence of writing $\nu(\mathbf{s}) = \mathbf{S}(\mathbf{s})'\eta, \quad \mathbf{s} \in D$, where η is a r dimensional vector with $\text{var}(\eta) = \mathbf{K}$ and $\nu(\cdot)$ is called a *spatial random effects* model. Now the Eq. (2.6) can be expressed as:

$$\mathbf{\Sigma} = \mathbf{S}\mathbf{K}\mathbf{S}^\top + \sigma^2\mathbf{V}. \tag{2.13}$$

Finally the kriging predictor (Eq. (2.7)) is:

$$\hat{Y}(\mathbf{s}_0) = \mathbf{x}(\mathbf{s}_0)^\top\hat{\alpha} + \mathbf{S}(\mathbf{s}_0)^\top\mathbf{K}\mathbf{S}^\top\mathbf{\Sigma}^{-1}(Z - \mathbf{X}\hat{\alpha}) \tag{2.14}$$

where

$$\mathbf{\Sigma}^{-1} = (\sigma^2\mathbf{V})^{-1} - (\sigma^2\mathbf{V})^{-1}\mathbf{S}\{\mathbf{K}^{-1} + \mathbf{S}^\top(\sigma^2\mathbf{V})^{-1}\}^{-1}\mathbf{S}^\top(\sigma^2\mathbf{V})^{-1}. \tag{2.15}$$

And the kriging standard error (Eq. (2.10)) is

$$\begin{aligned}\sigma_k(\mathbf{s}_0) = \{&\mathbf{S}(\mathbf{s}_0)^\top\mathbf{K}\mathbf{S}(\mathbf{s}_0) - \mathbf{S}(\mathbf{s}_0)^\top\mathbf{K}\mathbf{S}^\top\mathbf{\Sigma}^{-1}\mathbf{S}\mathbf{K}\mathbf{S}(\mathbf{s}_0) + \\ &(\mathbf{x}(\mathbf{s}_0) - \mathbf{X}^\top\Sigma^{-1}\mathbf{S}\mathbf{K}\mathbf{S}(\mathbf{s}_0))^\top(\mathbf{X}^\top\mathbf{\Sigma}^{-1}\mathbf{X})^{-1}(\mathbf{x}(\mathbf{s}_0) - \mathbf{X}^\top\Sigma^{-1}\mathbf{S}\mathbf{K}\mathbf{S}(\mathbf{s}_0))\}^{1/2}.\end{aligned} \tag{2.16}$$

For an extensive proof of the FRK refer [29]. One of the first applications was on total column ozone satellite data with $n = 173405$. The method is implemented in the FRK package [155] in the R statistical programming language. This method is applied to make inference on a hidden spatial process on the spatial domain of London in Section 6.2.

Even though Kriging can be applied to estimate revenues at new sites based on the neighbouring facilities, it ignores the underlying demand generating mechanisms in urban environments. Hence next section discusses a method that accounts for customer interactions in formulating the revenue generated at the facilities.

2.2 Spatial Interaction models

Spatial interaction models are determined to explain and predict patterns of interactions among entities over the geographic space [5]. These spatial interactions result from how entities in different locations make connections, choices, or demand/ supply decisions [129]. The entities can be people or firms, and the choices can include jobs, shopping facilities, school and sports activities. Spatial interaction frameworks applied for modelling flows such as migration [34, 135], commodities [5], retail sales [16, 95]. There is a long history of almost a century in formulating spatial interactions. I discuss important developments in literature, which was the motivation for my main study in this section.

Newton's Theory of Gravitation was prominent in the 19th century that applied to explain everything observable in the universe. Thus no surprise this shed light on describing certain types of peoples activity between entities in geographic space. Newton's Theory states that gravitational force F_{ij} between bodies i and j is:

$$F_{ij} = km_i m_j (d_{ij})^{-2} \tag{2.17}$$

where d_{ij} is the distance between the two masses m_i and m_j. Newton's model is translated to the conventional gravity models [22], to represent the interactions between origin i and destination j by F_{ij}, whereas m_i and m_j denotes the amount of activities produced at origin i and attracted to the destination j, and k is a constant of proportionality.

Reilly's law of retail gravitation [126] introduced in 1931 based on Newton's law of gravity acts as the reference model in the modern economic analysis of spatial interactions. He stated that the proportion of sales attracted by cities from an intermediate town is directly proportionate to the population in the two cities and inversely proportionate to the squares of the distance of the intermediate town. The general formulation is:

$$\frac{F_{ij}}{F_{ik}} = \left(\frac{m_j}{m_k}\right)\left(\frac{d_{ik}}{d_{ij}}\right)^2 \tag{2.18}$$

where F_{ij} and F_{ik} are the trade attracted from intermediate town O_i to cities

D_j and D_k respectively, m_j and m_k are the population of the two cities, and d_{ij} and d_{ik} are the distances from town to respective cities, and total distance between the two cities $d_{jk} = d_{ij} + d_{ik}$ (see Fig. 1.1).

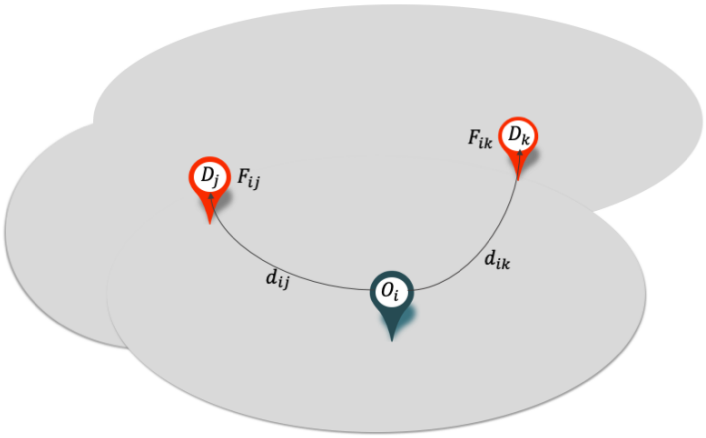


Figure 2.2: Illustration of trade attracted to cities from an intermediate town. Trade attracted to cities $F_{(\cdot)}$ from the intermediate town O_i and $d_{(\cdot)}$ represent the distance between locations.

Furthermore, the breaking point (d_b^*) indicates the point that one city dominates and beyond which the other city exercises retail trade influence. At the breaking point $F_{ij} = F_{ik}$ and the breaking point is:

$$d_b^* = \frac{d_{jk}}{1 + \sqrt{\frac{m_j}{m_k}}} \tag{2.19}$$

The above formulation is limited to evaluating competition between two retail centres concerning an intermediate town. However, with the growth in large shopping centres in cities in the '60s, assessing the competition within urban environments and between them was essential.

2.2.1 Huff model and extensions

Huff [78] in 1963 rose to the challenge by developing a probabilistic model shifting the focus from competitive retail firms to modelling the choices made by customers from alternative shopping options available in large urban environments. The *Huff model* utilises conceptual properties from the gravity model and focuses on understanding factors affecting the customer choices of shopping facilities and the choice process that leads to spatial behaviour. This leads to unveiling great insights and reliable estimates of the retail trade in urban areas.

The utility of a shopping centre to a customer depends on many factors. Huff argues that customers would travel extra distances or pick a store with

many items of their interest. Additionally, he considers the cost of travel to be inversely related to the shopping facility's choice. The probability p_{ij} of a customer travelling from zone O_i to shopping centre D_j is formulated as:

$$p_{ij} = \frac{w_j / t_{ij}^{\lambda}}{\sum_{j=1}^{n} w_j / t_{ij}^{\lambda}} \tag{2.20}$$

where w_j is the floor size of the shopping facility, t_{ij} denotes the cost of travel from location O_i to shopping centre D_j, and λ is the effect of travel time. The introduced *probabilistic rule* considers that customers patronise all facilities, and demand attracted by each facility is proportional to its attraction. This contrasts the *deterministic rule* that assumes customers buy only from one facility that attracts them the most [37, 77]. Applying the derived probability P_{ij} of a customer in zone O_i choosing shopping centre D_j, the expected revenue flow T_{ij} is:

$$T_{ij} = p_{ij} \times b_i \tag{2.21}$$

where b_i denotes the total budgeted spending or demand of customers in zone O_i. Another interpretation is that suppose b_i is the number of customers in zone O_i then T_{ij} provides the number of trips between O_i and D_j [129]. Suppose T_{ij} is expected revenue flow, then the total revenue r_j captured by the facility is:

$$r_j = \sum_i T_{ij} \tag{2.22}$$

Wilson [152] in 1969 suggested a variation on Huff formulation by stating that the probability a customer visiting a store is positively related to the facilities attraction and inversely related to a function of the distance between the customer and store locations d_{ij}. The exponential function was suggested to represent the distance decay:

$$p_{ij} = \frac{w_j^{\alpha} \exp(-\beta d_{ij})}{\sum_{j=1}^{n} w_j^{\alpha} \exp(-\beta d_{ij})} \tag{2.23}$$

where w_j now denotes the attraction of the facility and β is the distance decay parameter that assess the effect of distance. More generally the formulation is expressed as the *utility* u_{ij} gained by a customer i visiting the shopping facility in j:

$$p_{ij} = \frac{u_{ij}}{\sum_{j=1}^{n} u_{ij}}. \tag{2.24}$$

The model for estimating retail sales in shopping centres, the retail model

of Wilson [150] that he considered as the production constrained model is expressed closely related to the gravity model (Eq. (2.17)) as:

$$T_{ij} = b_i a_j w_j^{\alpha} \exp(-\beta d_{ij}) \tag{2.25}$$

$$a_j = \frac{1}{\sum_{j=1}^{n} w_j^{\alpha} \exp(-\beta d_{ij})} \tag{2.26}$$

and constraint on,

$$b_i = \sum_j T_{ij}. \tag{2.27}$$

Obtaining income level or demand at customer zones O_i is challenging in practice. Classical approach [150] is to represent as a product of mean expenditure and population in the customer region. But in recent studies, financial exposure in customer zones are estimated as a function of diverse factors, such as gender, age, household size, occupation, and income [16]. Similarly, instead of representing the facility's attraction with the floorspace, [105] proposed to express as a product of factors that are components of attractiveness and hence known as multiplicative competitive interaction (MCI) model. With the availability of data, both quantitative and qualitative factors such as offers, parking and age of the business are considered to represent the facility's attraction [16, 112]. In addition to the attractiveness of the stores, customers shop in areas with many facilities, such as shopping malls. This agglomeration effect is captured in the modified Huff model [95] by multiplying the attractiveness with the density of facilities around a store.

Furthermore, [69] suggested that competing shopping facilities may operate such that their floorspace cost balances the potential revenue generated from consumer disposable income discounted by the inconvenience of consumer travel. Thus it claimed that asymptotically in time, a deterministic equilibrium arises that is reflected by the most sustainable retail facility configuration:

$$\sum_i T_{ij} = kW_j \qquad \forall j \tag{2.28}$$

where W_j denotes the amount of floorspace and k is a constant.

One of the primary assumptions in the spatial interaction models is that the entire spending budget is distributed among the available shopping facilities, where $\sum_j p_{ij} = 1$. However, in practice, the entire demand of the customers are not satisfied due to lack of quality or availability. The decrease in spending at traditional retail facilities with the internet growth is even more prominent. This shortcoming is overcome by introducing a dummy facility [41] where the buying power attracted by this facility represents the lost demand.

The dummy facility is assumed to be located at the same distance d from all customers. The distance d signifies an appropriate distance the customers would travel to patronise a facility. Hence a more realistic formulation of the revenue flow from demand point O_i to the shopping facility D_j is;

$$T_{ij} = b_i \times \frac{w_j^{\alpha} \exp(-\beta d_{ij})}{\sum_{j=1}^{n} w_j^{\alpha} \exp(-\beta d_{ij}) + a \exp(-\beta d)}. \tag{2.29}$$

where dummy facilities attractiveness is denoted by a.

An essential step of operational modelling of urban systems is choosing a parameter estimation or calibration method. It is important to note that the equations (Eq. (2.25)) of spatial interactions are non-linear and cannot use well-developed linear statistical methods. One of the popularised approaches proposed by Wilson in 1969 [152] uses entropy-maximising principles. The final results are obtained by applying numerical methods such as Newton-Raphson scheme [7, 34]. Another approach is the log-linear analysis used commonly in migration contingency tables for estimating model parameters [123, 124]. Additionally, in literature, the parameters are commonly estimated by resorting to regression methods [8, 53, 95, 105]. Until recently, the focus of urban modelling was on the frequentist approaches that ignore the model and data uncertainty. Recently, model parameters have been estimated using the known spatial structure in a Bayesian manner by assuming that the facilities had reached a stochastic equilibrium status [46]. The Bayesian methods are limited to Markov Chain Monte Carlo MCMC [20, 46, 91] but does not scale up with large data. Calibration techniques related to Bayesian approach is discussed in more detail in Chapter 2.4.

One of the major challenges in calibrating spatial interaction model parameters is access to quality flow level data. In modelling migration flows, the researchers frequently use census data [34, 135]. However, it is more challenging to access transactional data from retail businesses to model retail flows of customers primarily due to its sensitivity. Thus model calibration and experiments are predominantly limited to synthetic data [2, 14, 105] or restrained to information collected by surveying households [62, 137].

Therefore, it is essential to adopt models for aggregate-level data representing the total sales generated by the customers at the stores to overcome the disaggregate level transactional data requirement. Furthermore, retail flow modelling is used to forecast market share or demand that is typically required to provide at the aggregate level of facilities. This would be more evident in the next section that illustrates the use of such models for identifying retail facility locations. While it is trivial in spatial interaction models forecasting aggregate level from the disaggregate model is challenging [129], the thesis address this in the modelling framework introduced in Chapter 3.

2.3 Competitive facility location problem

Facility location analysis is one of the major areas studied under operational research. The analysis depends on the industry that is of interest due to the differences in their motives in operating. Facilities such as warehouses, plants and distribution centres allocate customers by proximity that is discussed under proximity-based models [68] and p-median problem that determines to minimise the cost of transportation for satisfying the demand [67]. According to the deterministic proximity rule attractiveness of all the facilities are considered to be equal, and the total spending of a demand point is concentrated to one facility, "all or nothing" assumption [15]. In the context of locating emergency departments such as fire and ambulance services, the objective is to have the fewest number of sites so that all demand is covered within the stipulated maximum service response time, which is addressed with the location set covering problem [103, 140]. In contrast, *competitive facility location problems* emphasise industries such as retail businesses and commercial services, which consider competition among stores when choosing their sites [15, 40]. These businesses compete to attract customers buying power in a given area to capture market share, hence also known as *maximum capture problem*. This Chapter focuses more on the models developed to address the competitive facility location problem. Two major components can be identified in forming the model: estimation of market share and optimisation method.

2.3.1 Estimation of market share

The competitive facility location problem is focused on finding the best location to attract the highest market share in a particular region for a new retail or service facility(ies) while there are existing facilities. Hence it is important to calculate an accurate estimate of the market share or revenue the new facility can capture. The market share is calculated by allocating the customer budgeted spending among the facilities [43]. Multiple customer allocation rules are applied in the literature to model customer behaviour.

Gravity model

One approach to calculating the market share is to apply the probabilistic method introduced by Huff [78]. As discussed in the previous section, Huff type models [78, 152] suggest that customers located at specific demand points allocate their buying power among the competing facilities. The probability that customers patronise a facility is argued to be positively related to the attractiveness of the facility and inversely related to the cost of transport or distance. The total of the spending by customers at a facility provides the

revenue captured by the store (Eq. 2.22). Huff type models are proven to provide robust estimations market share [129] and are frequently applied in competitive facility location problems [15, 49, 75].

Let $\mathcal{I}$ be a set of customers each $i \in \mathcal{I}$ with a spending capacity of b_i in a competitive environment with an existing set of facilities J. Suppose the new facilities can be located at L and the market share M of the chosen locations of the new facilities $L^* \subset L$ is,

$$M = \sum_{i \in \mathcal{I}} b_i \frac{\sum_{l \in L^*} w_l^\alpha \exp(-\beta d_{il})}{\sum_{j \in J} w_j^\alpha \exp(-\beta d_{ij}) + \sum_{l \in L^*} w_l^\alpha \exp(-\beta d_{il})}. \tag{2.30}$$

Random Utility model

Another approach is the utility models [37] that formulates the expected customer satisfaction of each alternative facility with a utility function. This function is an aggregate measure of the trade-off between quality and distance to the facility. It is assumed that customers choose the facility with the highest utility. Thus the customers located at certain demand point has the same utility and select the same facility. In *random utility theory* [10, 54], the utility u_{ij} obtained by customer i choosing a facility comprises of deterministic v_{ij} and random term ϵ_{ij} such that,

$$u_{ij} = v_{ij} + \epsilon_{ij}. \tag{2.31}$$

The random term ϵ_{ij} represents the non observable attributes and the deterministic term is,

$$v_{ij} = w_j - \beta d_{ij} \tag{2.32}$$

where w_j average perceived quality of facility j and d_{ij} is the distance between i and j. Assuming ϵ_{ij} is identically and independently distributed, the model is referred to as the multinomial logit model and the probability a customer i choose facility j from the existing set of facilities J options is:

$$p_{ij} = \frac{\exp(v_{ij})}{\sum_{j \in J} \exp(v_{ij})}. \tag{2.33}$$

Then the market share captured by the new facilities is given by,

$$M = \sum_{i \in \mathcal{I}} b_i \frac{\sum_{l \in L^*} \exp(v_{il})}{\sum_{j \in J} \exp(v_{ij}) + \sum_{l \in L^*} \exp(v_{il})}. \tag{2.34}$$

2.3.2 Optimisation problem

The type of space determines the method of solving the optimisation problem. There are three types used in the literature. Continuous type problems consider anywhere in the plane as a potential location for the new facilities [38, 50], and network type of formulations regards any point on the network to be suitable for new facilities [132]. In contrast, the deterministic problem considers only a finite set of possible sites to locate the new facilities [10, 54]. I focus on the deterministic type of problem and also extend to search for optimal locations in competitive markets where the potential set of locations are initially not recognised.

The problem can be stated as; given a set of customers $i \in \mathcal{I}$ and their respective buying power or demand b_i, a set of existing competing facilities $j \in J$ and perceived utility of a customer i towards facility j is u_{ij}. Suppose the potential locations for new facilities $l \in L$ in an area are given, and then the problem is to locate q new facilities such that the expected market share captured by the new facilities is maximised. Let x_l be a binary variable that is set to 1 if and only if a new facility is to be located at l. The optimisation problem can be mathematically stated as:

$$\max_{x_l} \quad \sum_{i \in \mathcal{I}} b_i \frac{\sum_{l \in L} u_{il} x_l}{\sum_{j \in J} u_{ij} + \sum_{l \in L} u_{il} x_l} \tag{2.35}$$

$$\text{subject to:} \sum_{l \in L} x_l = q \tag{2.36a}$$

$$x_l \in \{0, 1\} \tag{2.36b}$$

The objective function is a sum of ratios (Eq. 2.35) and takes the form of a non-linear integer programming problem. This type of problem is known as combinatorial optimisation and is proven to be NP-hard [9]. There are multiple approaches employed to solve the problem such as branch-and-bound methods [9, 54], branch-and-cut [97] and heuristic methods [10]. I resort to IBM-ILOG CPLEX optimiser that provides constraint programming (CP) techniques to compute solutions for combinatorial optimisation problems [26].

CPLEX CP optimiser handles non-linear problems by applying a large set of arithmetic and logical constraints. The optimiser examines the model structure and observes constraint propagation to guide towards good solutions. There are multiple search types provided: depth-first search, restart search and multi-point search.

- **Depth-first search**

  The depth-first search method is a tree search algorithm that considers each decision variable as a branch in a search tree. The optimiser explores the branch's subtree and moves to the next section only after configuring a solution or proving that no solution exists in the current branch. Generally, this method is less efficient than restart search since it does not quickly recover from poor branching decisions.

- **Restart search**

  Restart search is the default search method in the optimiser that restarts the depth-first search after a certain number of failures to find an optimal solution. The number of failures between restarts can be adjusted from the model parameters.

- **Multi-point search**

  Multi-point search constructs a set of solutions and combines them to arrive at better solutions. Even though this method is diversified compared to other methods, it does not guarantee a solution's optimality or existence. The search runs until it reaches a point that the best solution found cannot be improved.

2.4 Bayesian Inference

As in all quantitative sciences, this thesis is concerned with data and learning the underlying processes in the economic system or source which generated data. The underlying data generating process is modelled with unknown parameters in mathematical formulations. There are two approaches in estimating these parameters: classical or frequentist and Bayesian methods. The frequentist approach focuses on finding the single best-fit parameter known as point estimates. This approach has drawbacks, such as the existence and the in-stability of the model due to data sensitivity. The Bayesian method accommodates for various shortcomings in frequentist modelling and provides many advantages such as a framework to quantify model or parameter uncertainty by assigning probability distributions to their possible values and applying prior knowledge to avoid overfitting [144]. Despite the drawbacks, much of the literature related to spatial interaction models are focused on the frequentist approach. Hence this thesis, resort to the Bayesian framework for the calibration of model parameters to make decisions under uncertainty.

In Bayesian modelling, two major types of uncertainties are modelled that are known as aleatoric and epistemic uncertainty [84]. Aleatoric explains the

noise inherent in the observed data. Epistemic accounts for the uncertainty in the model parameters also referred to as model uncertainty. The workflow of the Bayesian framework consists of three main steps. Initially capture the available knowledge about a particular parameter via a *prior distribution* which can be determined before collecting data. The *likelihood function* determines the information about the parameters available in the observed data. Finally, the Bayes' theorem, combining both prior and likelihood functions, forms the *posterior distribution*. Each step is explained in the next section.

The posterior distribution

Suppose the observed data denoted by $y \in \mathbb{R}^N$ and model parameters are denoted by $\theta \in \mathbb{R}^M$. The prior knowledge about θ are described by *prior distribution* $p(\theta)$. Then the likelihood $p(y|\theta)$ is expressed as the conditional distribution that describes how likely y to arise given θ. Finally, the *posterior distribution* provides the state of knowledge of θ parameters after observing y which is the conditional distribution obtain from the Baye's rule $p(\theta|y)$:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}. \tag{2.37}$$

The focus is to estimate the entire posterior distribution for the model parameters. The posterior distribution is summarised using the associated point estimates, such as posterior mean or median and a credible interval.

The unknown observable $\tilde{y}$ can be predicted from the same process. The distribution of $\tilde{y}$ that is conditional on the observed data y is called the *posterior predictive distribution*:

$$p(\tilde{y}|y) = \int p(\tilde{y}|y,\theta)p(\theta|y). \tag{2.38}$$

The likelihood function

The likelihood function explains any discrepancy between the model specifying underlying data generating process and observed data. The discrepancy is modelled as a random variable to express the uncertainty. The additive noise model is commonly used to express the observed values with a function f to model the underlying process and some random noise ϵ to represent the discrepancy between model and data:

$$y = f(\theta) + \epsilon. \tag{2.39}$$

The ϵ is assumed to have some distribution $\epsilon \sim \Upsilon$ and typical to assume

Gaussian distribution. The likelihood function is given by:

$$p(y|\theta) = \Upsilon(y - f(\theta)). \tag{2.40}$$

The prior distribution

The researcher is assumed to quantify the uncertainty about the parameters using their domain knowledge of the problem. Suppose the beliefs such as the values should be positive can be expressed with the statistical distribution called a *prior distribution* $p(\theta)$. The epistemic uncertainty is captured by placing prior distributions on the unknown parameters and observing their variation given the data [84]. The prior distributions can have different levels of informativeness. This is mainly classified into three levels: informative, weakly informative and diffuse [144].

The marginal likelihood

The posterior distribution is proportional to prior multiplied by likelihood and normalised constant known as *marginal likelihood* $p(y)$ or model evidence. The marginal likelihood is the distribution of the observed data and computed by marginalising out the parameters θ:

$$p(y) = \int p(y|\theta)p(\theta)d\theta \tag{2.41}$$

The integration in Eq. 2.41 often presents computational challenges. Mostly these integrals are high dimensional space and also analytically intractable. This was the primary reason to discard Bayesian statistics favouring frequentist approaches. However, Bayesian methods have gained popularity with the advancements in computer processing power and developments of approximate inference techniques. In the proceeding section, I discuss two widely used approximation techniques that I build upon in Chapter 3 : Variational Inference (VI) [81] and Markov Chain Monte Carlo (MCMC) [70].

2.4.1 Markov Chain Monte Carlo

The simulation process is possible using MCMC to calculate integrals involved in different forms of statistical inference [58]. This method is prominently applied in Bayesian inference to obtain posterior distribution using simulations [55, 57]. The MCMC method draws samples of θ from approximate distribution and improves those draws in a way it converges to the target posterior distribution $p(\theta|y)$ despite being high dimensional [56]. The sample θ values form the empirical estimates of the posterior distribution of interest and its related summary statistics such as mean, median and credible interval.

The MCMC method is a combination of two concepts: *Markov chains* are used to obtain parameter values from the posterior distribution, and *Monte Carlo integration* obtains samples to estimate the posterior distribution [144]. The Monte Carlo integration is a technique that uses computer simulations to estimate integrals by sampling from a certain distribution. Suppose θ is a random variable with a distribution of $p(\theta)$ then the expectation $\mathrm{E}(\theta)$ is:

$$\mathrm{E}(\theta) = \int \theta p(\theta) d\theta \tag{2.42}$$

which can be estimated by drawing n samples from the distribution $p(\theta)$:

$$\mathrm{E}(\theta) \approx \frac{1}{n} \sum_{i=1}^{n} \theta_i. \tag{2.43}$$

Kernel density estimation for the sampled values can produce the marginal posterior distribution of a certain parameter of interest.

The *Markov chain* is applied since it is impossible to sample directly and independently from the posterior distribution. In Markov chain simulation, a sequence of $\theta^1, \theta^2 \cdots$ is created by starting from θ^0 and draws θ^t from a transition distribution $T_t(\theta^t|\theta^{t-1})$ at each t that depends only on the last value θ^{t-1} [56]. The simulation is run sufficiently long such that the distribution of the current draws is close enough to the stationary distribution that specifies $p(\theta|y)$. Next, the standard Markov chain simulation methods are discussed: the Gibbs sampler, the Metropolis-Hastings algorithm, Hamiltonian Monte Carlo and the No-U-Turn Sample.

Metropolis and Metropolis-Hastings algorithms

Metropolis-Hastings (MH) [70] algorithms is a family of Markov chain methods that are useful for sampling from Bayesian posterior distributions. The *Metropolis algorithm* adapts a random walk that applies the acceptance/rejection rule to converge to the target distribution.

The *Metropolis-Hastings* algorithm generalises the Metropolis algorithm in two ways:

1. The proposal distribution need no longer be symmetric
2. Consequently to correct for the asymmetry the ratio is replaced by ratio of ratios

$$r = \frac{p(\theta^*|y)/q(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/q(\theta^{t-1}|\theta^*)} \tag{2.44}$$

The adoption of asymmetric rules allows the jumps to make a reasonable

Algorithm 1: The Metropolis algorithm

$\theta^0 \leftarrow$ Starting point drawn from initial distribution $p(\theta)$;

for $i = 1, 2 \cdots$ **do**

$\theta^* \leftarrow$ proposed point drawn from proposal distribtion at time t

$q_t(\theta^*|\theta^{t-1})$; `// Proposal distribtion must be symmetric`

$r = \dfrac{p(\theta^*|y)}{p(\theta^{t-1}|y)}$; `// Ratio of the densities`

$$\theta^t = \begin{cases} \theta^* & \text{with probability min(r,1)} \\ \theta^{t-1} & \text{otherwise} \end{cases}$$

end

distance that improves the speed of the random walk.

Gibbs sampler

Gibbs sampler [57] can be regarded as a special case of the Metropolis-Hastings algorithm. In the Gibbs sampler the parameter vector θ is subdivided in to d sub-vectors, $\theta = (\theta_1 \cdots \theta_d)$. At each iteration t the sample cycles through the d sub-vectors of θ drawing each subset condition on the value of all the others. At each iteration θ_j^t is sampled from the conditional distribution provided all the other components of θ:

$$\theta_j^t \sim p(\theta_j|\theta_{-j}^{t-1}, y), \tag{2.45}$$

where

$$\theta_{-j}^{t-1} = (\theta_1^t, \cdots, \theta_{j-1}^t, \theta_{j+1}^{t-1} \cdots \theta_d^{t-1}).$$

The Gibbs sampler is widely used since the ability of standard statistical models to sample from the conditional posterior distribution.

Hamiltonian Monte Carlo

Gibbs and Metropolis algorithms are inefficient in random walk behaviours, especially when the target distribution is high dimensional. In the Hamiltonian Monte Carlo (HMC) method [44, 98], a momentum term (ρ_j) is added for each component θ_j to help the chain move quickly to the target distribution. Both the terms are updated using the MH where the ρ_j gives distance and direction for θ_j to move through the space of θ rapidly. An independent distribution $p(\rho_j)$ is assumed on ρ_j giving the join distribution, $p(\theta, \rho|y) = p(\rho)p(\theta|y)$. The parameters are simulated from the joint distribution, but only interest in the

simulations of θ, and ρ is regarded as an auxiliary variable used to improve the algorithm to move faster in the parameter space. The simultaneous update of (θ, ρ) involves L 'leapfrog steps', and is scaled by a factor ϵ. The user has to specify these two parameters, and poor selection could lead to dramatically dropping HMC's efficiency.

No-U-Turn Sampler (NUTS)

No-U-Turn Sampler (NUTS) is an extension of HMC that overcomes the need to set the number of steps L [76]. A recursive algorithm runs the Hamiltonian simulation both forward and backwards in time to construct a set of possible values that span into a broader space of the target distribution. Additionally, a dual averaging approach is adopted to set the step size parameter ϵ automatically. The NUTS method is proven to run as efficiently as a well-tuned HMC method without requiring user interventions or the costly processes of finding optimal tuning parameters. Thus NUTS with dual averaging allows Bayesian analysts to apply HMC efficiently without much effort in hand-tuning the parameters. Hence adopt the NUTS method in calibrating the Bayesian model parameters in Chapter 3.

Assessing performance and computer software

The Markov chain needs to run many iterations before making reliable inferences on the posterior distribution. The period before the chain's convergence to the stationary distribution is regarded as the burn-in or warm-up period. Only the samples obtained after this stretch is used to summarise the posterior distribution and make final inferences. The trace plots are helpful for visually exploring the parameters' behaviour over the iterations as illustrated in Fig. 2.3. These plots are used to determine when the Markov chain has reached its stationary distribution. The formal procedure to decide the convergence of the Markov chain is the R statistic that defines the ratio of within-chain to between-chain variability [21]. R-value close to one indicates that the Markov chain has converged to its stationary distribution, and the samples can be regarded as from the posterior distribution.

There are software packages that have implemented Bayesian analysis, and Stan is an open-source package developed in C++. The Rstan package [134] with the R interface is adopted to execute the MCMC algorithms and to generate diagnostics for assessing performance.

2.4.2 Variational Inference

MCMC has been the dominant approach for the approximate inference required to compute complex Bayesian models for many decades [70]. However,

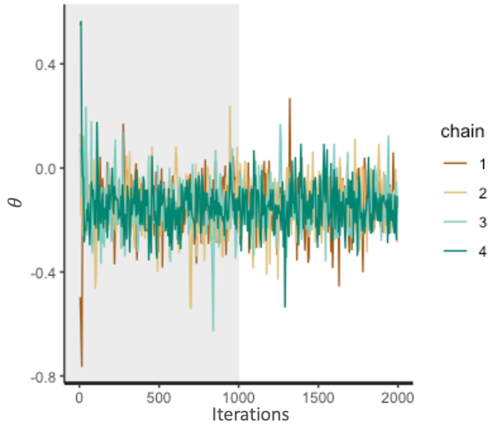


Figure 2.3: Trace plot illustrates the iteration number against a parameter value for four independent chains of the MCMC. The shaded section represents the warm-up phase which is omitted in creating the posterior distribution.

this method does not perform well with large scale datasets or complex models. In such cases, similar to the problem that is interested in solving, spanning into large scale spatial data, variational inference (VI) provides a better approach to approximate Bayesian inference. In contrast to sampling techniques used in MCMC, the VI approach solves an optimisation problem.

The objective of variational inference is to approximate a posterior distribution or conditional density of unknown parameters or latent variables given the observations, $p(\theta|y)$. A family Q of densities is specified over θ, where each $q(\theta) \in Q$ is considered a candidate approximation of the posterior distribution of interest. The Kullback–Leibler (KL) divergence [86], a statistical distance measure between two densities, is employed to identify the best candidate $q^*(\theta)$ closest to the exact posterior. The inference is now obtained by solving an optimisation problem:

$$q^*(\theta) = \min_{q(\theta)\in Q} \mathrm{KL}\left(q(\theta)||p(\theta|y)\right) \tag{2.46}$$

The (KL) divergence is given by:

$$\mathrm{KL}\left(q(\theta)||p(\theta|y)\right) = \mathbb{E}[\log q(\theta)] - \mathbb{E}[\log p(\theta|y)], \tag{2.47}$$

where all the expectation are in respect to $q(\theta)$. Expanding the conditional density leads to,

$$\mathrm{KL}\left(q(\theta)||p(\theta|y)\right) = \mathbb{E}[\log q(\theta)] - \mathbb{E}[\log p(\theta, y)] + \log p(y). \tag{2.48}$$

This shows the dependency on $p(y)$ that possesses computational challenges as discussed in Eq. 2.41 which compelled us to adopt approximate inference

in solving the posterior distribution. Hence instead, optimise an alternative objective function by disregarding the constant $p(y)$ with respect to $q(\theta)$. This function is called the evidence lower bound, ELBO :

$$\text{ELBO}(q) = \mathbb{E}[\log p(\theta, y)] - \mathbb{E}[\log q(\theta)], \tag{2.49}$$

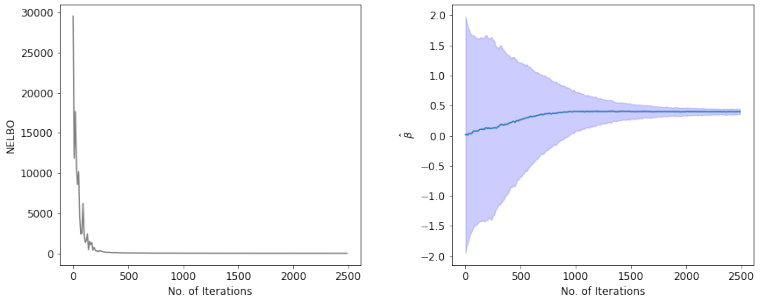
which is the negative KL divergence of Eq. 2.48 without the constant $p(y)$. Thus minimising the KL is equivalent to maximising ELBO. The ELBO can be expressed as a sum of the expected log likelihood of the data and the KL divergence between $q(\theta)$ and prior $p(\theta)$:

$$\begin{aligned} \text{ELBO}(q) &= \mathbb{E}[\log p(\theta)] + \mathbb{E}[\log p(y|\theta)] - \mathbb{E}[\log q(\theta)] && (2.50) \\ &= \mathbb{E}[\log p(y|\theta)] - \text{KL}(q(\theta)||p(\theta)). && (2.51) \end{aligned}$$

The reasoning for the term ELBO can be explained by combining Eqs. 2.48 – 2.49:

$$\log p(y) = \text{KL}(q(\theta)||p(\theta|y)) + \text{ELBO}(q) \tag{2.52}$$

and since $\text{KL}(\cdot) \geq 0$ [82], ELBO is lower bound to $\log p(y)$ for any $q(\theta)$, $\log p(y) \geq \text{ELBO}(q)$ [18]. The negative ELBO is minimised where the convergence over iterations are illustrated for ELBO and an unknown parameter in Fig. 2.4.



(a) (b)

Figure 2.4: Illustrates the number of iteration against (a) negative ELBO (b) a parameter value

The variational family Q is described in various ways. The class of distributions primarily determines the complexity of the optimisation. Hence this thesis focus on a standard approach, *mean-field variational family* that considers θ to be mutually independent [56]:

$$q(\theta) = \prod_{i=1}^{j} q_j(\theta_j), \tag{2.53}$$

where each unknown parameter is governed by its own variational density $q_j(\theta_j)$. Each of these densities can take any form appropriate to the corresponding random variable.

The variational algorithm is formulated by expanding Eq. 2.50 that maximises the ELBO over variational parameters. Then applies an optimisation procedure by computing gradients. A limited set of variational families, such as conjugate exponential family models, can derive a closed-form [59]. Other models require analytic computations of various expectations that can be tedious. To overcome this, the *black-box variational inference* approach that calculates expectations with Monte Carlo samples from the variational distribution is adopted [121].

The variational inference formulation is translated to optimise ELBO in Python programming software [145]. More specifically, the main model is built on TensorFlow, which is an open-source platform for machine learning [1].

2.4.3 Comparison between VI and MCMC

This section compares and summarises the approximation techniques for Bayesian analysis discussed. The MCMC method simulates from densities to form empirical estimates, and VI is a technique that approximates densities. MCMC methods guarantee exact samples from the target densities, whereas VI only find a close distribution to the target. VI tends to underestimate the posterior variance compared to MCMC [18] and is illustrated in Fig. 2.5 . Thus in cases that require precise inferences, it should adopt MCMC methods. Based on the problem, underestimation of variance may be acceptable.

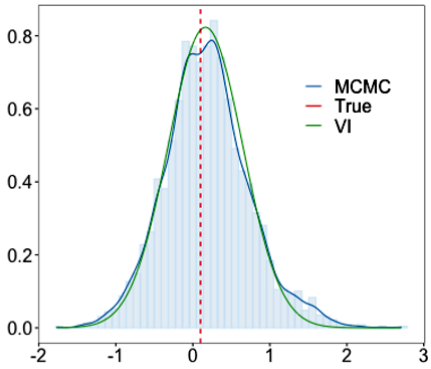


Figure 2.5: Comparison between the approximate distributions that results from MCMC and VI. Histogram shows the empirical distribution from MCMC and a kernel density is fitted. Approximate density of the posterior from VI is generated using the estimated distribution parameters.

MCMC is computationally intensive and slower than VI technique that explores the advantages of using various optimisation processes. Hence VI is

more suitable for making inferences with large datasets, whereas MCMC is more appropriate for small scale problems. I compare approximate posterior distributions in Chapter 3, using a simulation study to explore the variations in the two methods. Based on the performance, the VI technique is adopted in addressing real-world problems with large datasets in Chapter 6.

2.5 Summary

The scope of this Chapter has provided the reader with the essential background knowledge required to fully understand the remainder of this thesis: (1) described a popular spatial model for spatial prediction, Kriging and its extended variation of Fixed ranked Kriging; (2) put forth a history of spatial interaction models and the theory behind them; (3) discussed the competitive facility location problem and its main components; (4) comprehensive introduction to Bayesian inference and approximation techniques - VI and MCMC.

Next chapter introduce the Bayesian spatial interaction model with a scalable variational inference framework that, while being significantly faster than competing Markov Chain Monte Carlo inference schemes. Furthermmore. demonstrate the benefits of BSIM in various synthetic settings characterised by an increasing number of stores and customers.

Chapter 3

A Bayesian spatial interaction model for estimating revenue and demand at business facilities

3.1 Introduction

Understanding the interaction between business facilities and consumer preferences is a prime factor of success for industries such as retail, healthcare and hospitality. Therefore, accurate predictions of potential sales at business locations are becoming crucial for planning and decision-making in the current ecosystem. Indeed, the continuous growth in e-commerce [110] is threatening the existence of traditional retail stores. I propose a Bayesian statistical methodology that, by capturing the relationship between attractiveness of the facility, distance between a business location and its customers, and demand in terms of buying power, allows to make probabilistic forecasts about potential revenue at a business facility while quantifying the uncertainty in these estimates.

As discussed in Chapter 2.2, one of the earliest mathematical models of customer behaviours when choosing shopping facilities is known as Law of Retail Gravitation [125], which was inspired by the Newtonian gravity model and formulated a customer's choice between two facilities as a function of their attractiveness and distances. Huff [78] subsequently extended this model to consider multiple facilities while providing a probabilistic interpretation for the spatial interactions between customers and facilities. In the following years, the Huff model [78] was improved by replacing the single attractiveness term determined by floorspace with a composite index of a set of attributes

at the facility, including economic and structural factors [32, 105]. Most of the literature estimates the parameters of spatial interaction models by resorting to regression methods [8, 53, 95, 105] or by maximising the entropy with respect to some constraints [52, 151]. More recently, computationally intensive Markov Chain Monte Carlo (MCMC) schemes have been proposed as an alternative inference method within the Bayesian framework for modelling origin-destination flows but do not offer capabilities in estimating total revenue or demand generated at the destination [25, 46, 90].

Inspired by the literature on gravity models, I have developed a Bayesian spatial interaction model, henceforth named BSIM, which provides probabilistic predictions about revenues generated at business facilities given their features and the potential customers' characteristics in a specified region in space. The probability of a customer visiting each facility in a region is modelled through Gaussian densities in geographic space. Specifically, each density is centred on a facility with a variance that is further determined by its attractiveness which in turn is modelled as a function of internal and external characteristics (e.g. floorspace, distance to public transport access points) and customer perspective (e.g. customer rating). The revenues for each facility are then obtained by combining the probability of a customer visit with a proxy of the individuals buying power, which is assumed to be a function of their socio-demographic characteristics. I adopt a Bayesian approach that enables to adequately account for the uncertainty associated with the customer interactions with the facilities. My framework not only gives accurate predictions but produces interpretable results that can support experts' decision-making processes. Moreover, this approach allows us to infer quantities at the business facility or customer level, such as revenue flow from customers to businesses. In BSIM, the posterior distributions of interest are intractable, and their approximation poses significant computational challenges. This issue is addressed by resorting to variational inference while also comparing with MCMC approximation. I demonstrate how the variational scheme is significantly faster compared to MCMC used in the literature while providing comparable results in terms of parameter identification and uncertainty quantification.

My main contributions from this chapter are: (a) develop a Bayesian spatial interaction model (BSIM) that can be used to make probabilistic predictions of revenues or demand generated at business facilities; (b) introduce a probabilistic method to formulate the relationship between distance and attractiveness of facilities jointly, using a facility-specific probability distribution; (c) propose a scalable variational inference and demonstrate its benefits compared to MCMC methods in a variety of experimental settings.

3.2 Methodology

Consider a regression problem for a given dataset $\mathcal{D} = \{\mathbf{x}_s, y_s\}_{s=1}^{S} \cup \{\mathbf{v}_n\}_{n=1}^{N}$, where $\mathbf{x}_s \in \mathbb{R}^D$ represents the s-th store[1] features and $y_s \in \mathbb{R}$ gives the revenue for the s-th store and $\mathbf{v}_n \in \mathbb{R}^P$ represents the features of the n-th customer in a bounded region τ. Each feature vector $\mathbf{x}_s^\top = [\mathbf{l}_s^\top, \boldsymbol{\phi}_s^\top]$ includes the store location, which is denoted by $\mathbf{l}_s \in \mathbb{R}^2$, and additional store characteristics denoted by $\boldsymbol{\phi}_s \in \mathbb{R}^{D-2}$, e.g. floor size. For notational convenience $S \times (D-2)$ matrix denotes all stores characteristics by $\boldsymbol{\Phi}$. Suppose there exists N customers within τ where $\mathbf{v}_n^\top = [\mathbf{m}_n^\top, \mathbf{w}_n^\top]$ includes the customer location, which is denoted by $\mathbf{m}_n \in \mathbb{R}^2$ and its characteristics denote by $\mathbf{w}_n \in \mathbb{R}^{P-2}$ such as income level.

3.2.1 Model Formulation

The proposed Bayesian Spatial Interaction Model (BSIM) is characterised by S Gaussian distributions, one for each store, which are uncorrelated a priori. Each Gaussian distribution is centered on a store's location $\boldsymbol{\mu}_s = \mathbf{l}_s$ and has a diagonal covariance matrix $\boldsymbol{\Sigma}_s = \sigma_s^2 \mathbf{I}$, henceforth $\mathcal{N}(\boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$. The variance σ_s^2 captures level of "attraction" of a customer to a store. I propose two different alternative forms for variance. In the first model σ_s^2 is written as a function of store specific coefficient $\upsilon_s \in \mathbb{R}$ that is:

$$\sigma_s^2 = \exp(\upsilon_s), \tag{3.1}$$

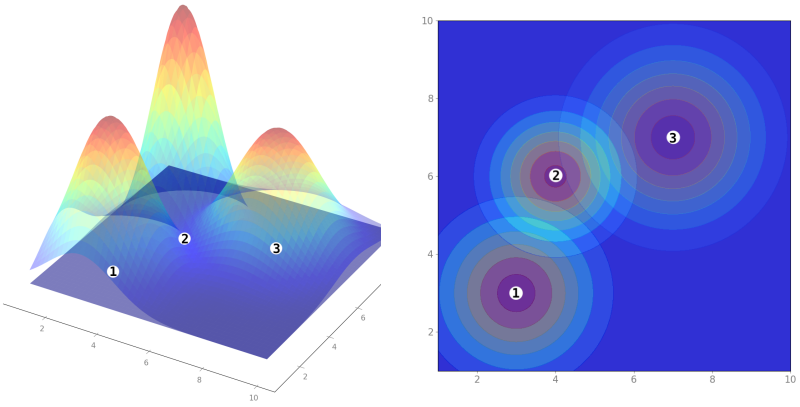
In the second, the specification is improved by denoting υ_s as a function of store characteristics:

$$\upsilon_s = \boldsymbol{\lambda}^\top \boldsymbol{\phi}_s + \varepsilon_s, \tag{3.2}$$

where $\boldsymbol{\lambda} \in \mathbb{R}^{D-2}$ represents shared coefficients across the stores, and ε_s denotes the non-observable store characteristics. The probability density function (PDF) of the Gaussian distribution centred on store location $\mathbf{l}_s$ evaluated at location $\mathbf{m}_n$ which allows us to capture the likelihood for the n-th customer to visit the s-th store based on their distance and on the store characteristics. For illustration purposes, consider three stores where each has a Gaussian distribution centred on the store, as shown in Fig. 3.1.

Irrespective of the store's attractiveness, customer behaviour is not affected after a certain maximum distance to the store, known as "consideration set" in marketing. Therefore, truncate the Gaussian distributions in BSIM and force

[1]The the rest of the model is presented in relation to the specific instantiation where a business location is a store, but this can be extended to other business facilities.



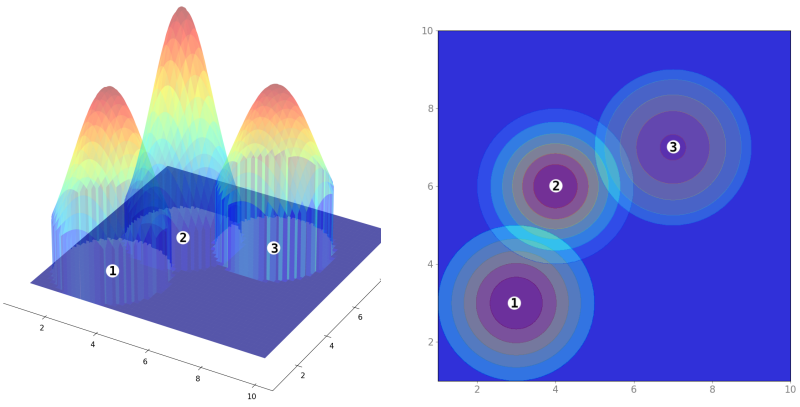
(a) (b)

Figure 3.1: Illustration of the PDF of the Gaussian distribution centered on three sample Stores: (a) 3D visualisation; (b) 2D visualisation. The white dots indicate the store location and the numbers are used to identify the respective stores on 3D and 2D visualisations.

their densities to be zero beyond a given radius d_T from the store location. The truncated Gaussian PDF denoted by $f(d_{ns})$ is given by:

$$f(d_{ns}) = \begin{cases} \dfrac{\exp\left(-d_{ns}^2/2\sigma_s^2\right)}{2\pi\sigma_s^2\left(1-\exp\left(-d_T^2/2\sigma_s^2\right)\right)}, & 0 \leq d_{ns} \leq d_T, \\ 0, & \text{otherwise}, \end{cases} \tag{3.3}$$

where d_{ns} denotes the Euclidean distance between the store and customer $d_{ns} = ||\mathbf{m}_n - \mathbf{l}_s||_2$. Fig. 3.2 demonstrates the truncated Gaussian densities corresponding to the distributions shown in Fig. 3.1.



(a) (b)

Figure 3.2: Illustration of the Truncated Gaussian centered on three sample Stores: (a) 3D visualisation; (b) 2D visualisation. The white dots indicate the store location. There is a hard border around the distributions beyond which the PDF is equal to zero.

Given the truncated Gaussian distributions, the probability p_{ns} of a customer visiting the s-th store is defined as:

$$p_{ns} = \frac{f(d_{ns})}{\sum_{j=1}^{S} f(d_{nj})}. \tag{3.4}$$

Note that I normalise the PDF calculated for the customer with respect to the store by the total PDF respect to all the stores within the consideration set to arrive at a value that falls in the interval of [0, 1]. Thus it is assumed that every customer chooses at least one store in their consideration set, but this can be relaxed by adding pseudo stores to account for unsatisfied demand or unobserved data. The value of p_{ns} captures the level of competition in the region τ for a specific type of store. For instance, p_{ns} will be lower in competitive markets or areas while it will take higher values in non-competitive settings. This is illustrated in Fig. 3.3 with respect to the non-truncated and truncated Gaussian distributions.

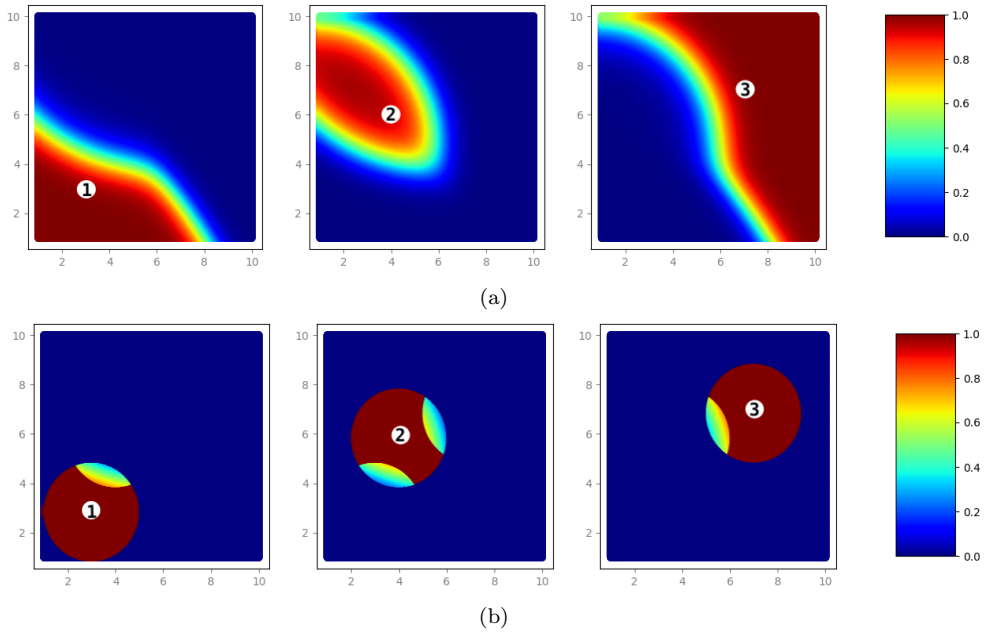


Figure 3.3: Illustration of the probability of customers visiting a store p_{ns}: (a) with none truncated Gaussian distribution; (b) with truncated Gaussian distribution. This is an indication of the competition in the area. The white dots indicate the store location, and the numbers are used to identify the respective stores on (a) and (b) plots.

The consumption function in economics determines the relationship between consumer spending and the various factors [102]. To model the amount budgeted by each customer for spending b_n, I propose a linear function $g(\cdot)$ which takes

input $\mathbf{w}_n$ representing the $P-2$ customer characteristics:

$$b_n = g(\mathbf{w}_n) = \boldsymbol{\beta}^\top \mathbf{w}_n, \tag{3.5}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{P-2}$. This leads to the conventional Spatial interaction system [46, 78, 150]. Thus expenditure flow from customer n to store s:

$$r_{ns} = b_n \times p_{ns}, \tag{3.6}$$

where the amount each customer budgeted to spend b_n is weighed by the probability to visit the s-th store. The total revenue for the s-th store is:

$$r_s = \sum_{n=1}^{N} b_n p_{ns} = \sum_{n=1}^{N} \boldsymbol{\beta}^\top \mathbf{w}_n \frac{f(d_{ns})}{\sum_{j=1}^{S} f(d_{ns})}. \tag{3.7}$$

Henceforth I derive the model for the case where the store variance is a function of its features (Eq. (3.2)), since the limiting case where the store variance is store specific coefficient (Eq. (3.1)) is a trivial extension by setting λ to zero.

Likelihood function: The likelihood of the observed stores' revenue $\mathbf{Y} = \{y_1, \ldots, y_S\}$ is defined as:

$$p(\mathbf{Y}|\boldsymbol{\beta}, \lambda, \varepsilon, \gamma) = \prod_{s=1}^{S} \mathcal{N}\left(y_s | \sum_{n=1}^{N} \boldsymbol{\beta}^\top \mathbf{w}_n \frac{f(d_{ns})}{\sum_{j=1}^{S} f(d_{ns})}, \gamma^{-1}\right), \tag{3.8}$$

where the model assumes constant noise precision (γ^{-1}) for the Gaussian.

Prior Distributions: Prior distributions are assigned to all model parameters. First, a hierarchical prior distribution is defined for $\boldsymbol{\beta}$, which is assumed to be a Gaussian with mean μ_β and covariance $\alpha^{-1}\mathbf{I}$:

$$p(\boldsymbol{\beta}|\alpha) = \mathcal{N}(\boldsymbol{\beta}; \mu_\beta, \alpha^{-1}\mathbf{I}),$$

Following the standard practices, a Gamma prior distribution is introduced with shape $\omega_1 > 0$ and scale $\omega_2 > 0$ for the hyper-parameter α:

$$p(\alpha) = \mathrm{Gam}(\alpha; \omega_1, \omega_2)$$

Similarly, a Gamma prior distribution with shape ρ_1 and scale ρ_2 is assumed for the likelihood precision parameter γ:

$$p(\gamma) = \mathrm{Gam}(\gamma; \rho_1, \rho_2),$$

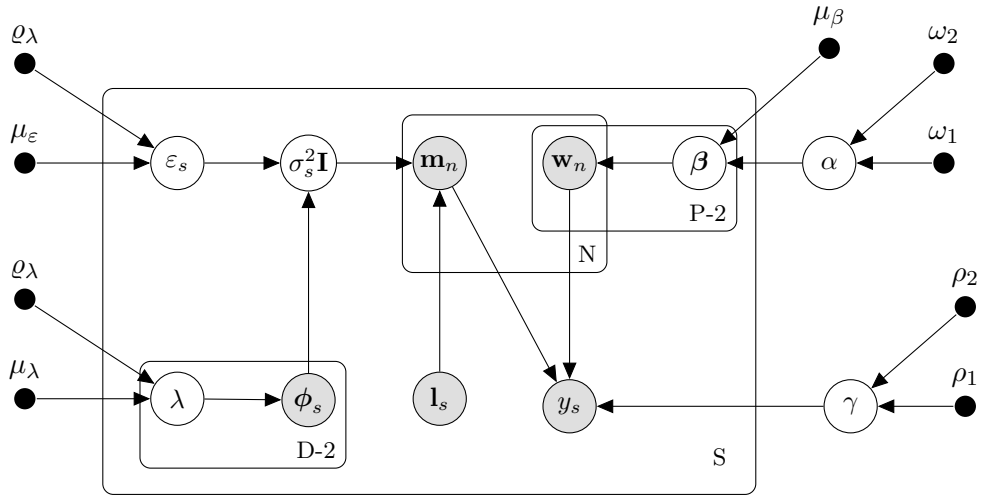


Figure 3.4: Plate diagram for the graphical representation for the BSIM. This express the spatial interaction between S number of stores with each store revenue y_s, located at $\mathbf{l}_s$ with store features $\boldsymbol{\phi}_s$ and N number of customers located at $\mathbf{m}_n$ with P-2 characteristics $\mathbf{w}_n$. Gaussian distributions are used as priors for $\boldsymbol{\beta}, \lambda, \varepsilon$ and Gamma distributions for γ, α. The diagram represents random variables with circles (◯), known values with grey filled circles (◯) while black filled circles (●) indicate fixed parameters of prior and hyper-prior distributions, edges denote possible dependence, and plates denote replication.

Finally, the following Gaussian prior distributions are selected for λ and ε,

$$p(\lambda) = \mathcal{N}(\lambda; \mu_\lambda, \varrho_\lambda \mathbf{I})$$
$$p(\varepsilon) = \mathcal{N}(\varepsilon; \mu_\varepsilon, \varrho_\varepsilon \mathbf{I}).$$

Posterior Distribution: The full vector of model parameters is denoted by $\boldsymbol{\Theta} = \{\boldsymbol{\beta}, \lambda, \alpha, \varepsilon, \gamma\}$. Posterior probability given by:

$$p(\boldsymbol{\Theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})}{\int p(\mathcal{D}|\boldsymbol{\Theta})p(\boldsymbol{\Theta})\mathrm{d}\boldsymbol{\Theta}} \tag{3.9}$$

where the marginal density takes the form:

$$p(\mathcal{D}) = \int \cdots \int p(\mathcal{D}|\boldsymbol{\beta}, \lambda, \gamma)p(\boldsymbol{\beta}|\alpha)p(\alpha)p(\lambda)p(\boldsymbol{\varepsilon})p(\gamma)\,\mathrm{d}\boldsymbol{\beta}\,\mathrm{d}\alpha\,\mathrm{d}\lambda\,\mathrm{d}\boldsymbol{\varepsilon}\,\mathrm{d}\gamma. \tag{3.10}$$

3.2.2 Inference

Our goal is to estimate the posterior distribution over all parameters given the data i.e. $p(\boldsymbol{\Theta}|\mathcal{D})$. Since marginal density is analytically intractable (Eq. (3.10)), I resort to approximate inference by employing two commonly used methods: Variational Inference (VI) [81] and Markov Chain Monte Carlo (MCMC) [70].

Variational Inference

VI is a powerful method to approximate intractable integrals, whereas in contrast to MCMC, it tends to be much faster because it rests on optimisation instead of sampling [18]. VI first posits a family of densities and then finds the member of that family, which is closest to the posterior, by minimising the Kullback-Leiber (KL) divergence. Because the KL divergence cannot be directly calculated, alternatively, maximise evidence lower bound, $\mathcal{L}_{\text{elbo}}$ that is equivalent to minimizing the KL divergence. A more detailed explanation is provided in Chapter 2.4.

Variational Distributions: The mean-field approximation is adopted and assume a fully factorized variational distribution [17]:

$$q(\boldsymbol{\beta}, \alpha, \gamma, \lambda, \varepsilon) = q(\boldsymbol{\beta})q(\alpha)q(\gamma)q(\lambda)q(\varepsilon), \tag{3.11}$$

with

$$q(\boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\beta}; \hat{\mu_\beta}, \boldsymbol{\Omega}) \tag{3.12}$$

$$q(\alpha) = \text{Gam}(\alpha; \hat{\omega_1}, \hat{\omega_2}), \tag{3.13}$$
$$q(\gamma) = \text{Gam}(\gamma; \hat{\rho}_1, \hat{\rho}_2), \tag{3.14}$$
$$q(\lambda) = \mathcal{N}(\lambda; \hat{\mu_\lambda}, \mathbf{K}_\lambda), \tag{3.15}$$
$$q(\varepsilon) = \mathcal{N}(\varepsilon; \hat{\mu_\varepsilon}, \mathbf{K}_\varepsilon), \tag{3.16}$$

where $\boldsymbol{\nu} = \{\hat{\mu_\beta}, \boldsymbol{\Omega}, \hat{\omega_1}, \hat{\omega_2}, \hat{\rho}_1, \hat{\rho}_2, \hat{\mu_\lambda}, \mathbf{K}_\lambda, \hat{\mu_\varepsilon}, \mathbf{K}_\varepsilon\}$ are the variational parameters which are optimized within the algorithm. Eqs. (3.12)–(3.16) define our approximate posterior. With this, details of the variational objective function is given, i.e. $ELBO$, which aims to maximise with respect to $\boldsymbol{\nu}$.

Evidence Lower Bound: Following the standard variational inference, ELBO can be written as a combination of expected log likelihood ($\mathcal{L}_{\text{ell}}$) and KL-divergence term ($\mathcal{L}_{\text{kl}}$):

$$\mathcal{L}_{\text{elbo}}(\boldsymbol{\nu}) = \mathcal{L}_{\text{ell}}(\boldsymbol{\nu}) - \mathcal{L}_{\text{kl}}(\boldsymbol{\nu}). \tag{3.17}$$

the expected log likelihood term can be written as

$$\begin{aligned}\mathcal{L}_{\text{ell}} &= \mathbb{E}_{\boldsymbol{\beta},\gamma,\lambda,\varepsilon}[\ln p(Y|\boldsymbol{\beta},\gamma,\lambda,\varepsilon)] \\ &= -\frac{S}{2}\ln 2\pi + \frac{S}{2}(\psi(\hat{\rho}_1) - \ln\hat{\rho}_2) - \\ &\quad \frac{1}{2}\frac{\hat{\rho}_1}{\hat{\rho}_2}\mathbb{E}_{\boldsymbol{\beta},\gamma,\lambda,\varepsilon}\left[\frac{\gamma}{2}\sum_{s=1}^{S}\left(y_s - \boldsymbol{\beta}^\top\sum_{n=1}^{N}\mathbf{w}_n\frac{f(d_{ns})}{\sum_{j=1}^{S}f(d_{ns})}\right)^2\right]\end{aligned} \tag{3.18}$$

the KL-Divergence Term is expanded and simplified as:

$$\begin{aligned}\mathcal{L}_{\text{kl}} &= \mathbb{E}[\ln p(\boldsymbol{\Theta})] - \mathbb{E}[\ln q(\boldsymbol{\Theta})] \\ &= \mathbb{E}_{\boldsymbol{\beta},\alpha}[\ln(p(\boldsymbol{\beta}|\alpha)] + \mathbb{E}_\alpha[\ln p(\alpha)] + \mathbb{E}_\lambda[\ln p(\lambda)] + \mathbb{E}_\varepsilon[\ln p(\varepsilon)] + \mathbb{E}_\gamma[\ln p(\gamma)] - \\ &\quad \mathbb{E}_{\boldsymbol{\beta}}[\ln q(\boldsymbol{\beta})] - \mathbb{E}_\alpha[\ln q(\alpha)] - \mathbb{E}_\lambda[\ln q(\lambda)] - \mathbb{E}_\varepsilon[\ln q(\varepsilon)] - \mathbb{E}_\gamma[\ln q(\gamma)],\end{aligned} \tag{3.19}$$

where each term is expanded and simplified as:

$$\begin{aligned}\mathbb{E}_{\beta,\alpha}[\ln(p(\beta|\mu_\beta,\alpha)] = &-\frac{P}{2}\ln(2\pi) + \frac{P}{2}(\psi(\hat{\omega_1}) - \ln\hat{\omega_2}) \\ &- \frac{\hat{\omega_1}}{2\hat{\omega_2}}\left(tr(\boldsymbol{\Omega}) + \hat{\mu_\beta}^\top\hat{\mu_\beta} - 2\mu_\beta^\top\hat{\mu_\beta} + \mu_\beta^\top\mu_\beta\right)\end{aligned} \tag{3.20}$$

where ψ is the digamma function.

$$\mathbb{E}_\alpha[\ln p(\alpha)] = \omega_1\ln\omega_2 + (\omega_1 - 1)(\psi(\hat{\omega}_1) - \ln\hat{\omega}_2) - \omega_2\frac{\omega_1}{\hat{\omega}_2} - \ln\Gamma(\omega_1) \tag{3.21}$$

$$\mathbb{E}_\gamma[\ln p(\gamma)] = \rho_1\ln\rho_2 + (\rho_1 - 1)(\psi(\hat{\rho}_1) - \ln\hat{\rho}_2) - \rho_2\frac{\hat{\rho}_1}{\hat{\rho}_2} - \ln\Gamma(\rho_1) \tag{3.22}$$

$$\mathbb{E}_\lambda[\ln(p(\lambda)] = -\frac{m}{2}\ln(2\pi\varrho) - \frac{1}{2\varrho}(tr(\mathbf{K}_\lambda) + \hat{\mu_\lambda}^\top\hat{\mu_\lambda} - 2\mu_\lambda^\top j_\lambda + \mu_\lambda^\top\mu_\lambda) \tag{3.23}$$

$$\mathbb{E}_\beta[\ln q(\beta)] = -\frac{1}{2}\ln|\boldsymbol{\Omega}| - \frac{P}{2}(\ln(2\pi) + 1) \tag{3.24}$$

$$\mathbb{E}_\gamma[\ln q(\gamma)] = (\hat{\rho}_1 - 1)\psi(\hat{\rho}_1) + \ln\hat{\rho}_2 - \hat{\rho}_1 - \ln\Gamma(\hat{\rho}_1) \tag{3.25}$$

$$\mathbb{E}_\alpha[\ln q(\alpha)] = (\hat{\omega}_1 - 1)\psi(\hat{\omega}_1) + \ln\omega_2 - \hat{\omega}_1 - \ln\Gamma(\hat{\omega}_1) \tag{3.26}$$

$$\mathbb{E}_\lambda[\ln q(\lambda)] = -\frac{1}{2}\ln|\mathbf{K}_\lambda| - \frac{m}{2}(\ln(2\pi) + 1) \tag{3.27}$$

$\mathcal{L}_{\text{elbo}}(\boldsymbol{\nu})$ is not computable in analytically closed forms and remains intractable. Hence I resort to Black Box variational inference method where the gradient is computed from the Monte Carlo samples from the variational distributions [121]. The algorithm is implemented in Tensorflow 2 [1] in Python 3.

Markov Chain Monte Carlo

In order to compare the estimations, MCMC is employed, which has been the dominant paradigm for approximate inference for decades. First, a Markov chain on $\boldsymbol{\Theta}$ is constructed whose stationary distribution is the posterior $p(\boldsymbol{\Theta}|\mathcal{D})$. Then samples are collected from the stationary distribution by sampling from the Markov chain. Finally, the collected samples are used to approximate the posterior with an empirical estimate. MCMC methods ensure producing exact samples from the target density but tend to be computationally intensive [127]. When the datasets are large, MCMC becomes slower and computationally expensive to form inferences. I use open-source software, Stan, which is a C++ library for Bayesian modelling, with the R interface to compile results [134]. The No-U-Turn sampling method is adopted, an extension to the Hamiltonian Monte Carlo algorithm for the experiments [76]. The MCMC methods are discussed in Chapter 2.4.

3.2.3 Edge Correction

Stores on the edge of the study area τ cannot be evaluated without a certain bias because the model cannot capture the contribution from customers living outside τ. To overcome this, the revenues of the stores $\{y_s\}_{s=1}^{S}$ are adjusted, and this is carried out before fitting the model. Following a similar approach to the model, I assume a Gaussian centred on the store and calculate the area under the curve (AUC) $\mathcal{A}$, which intersects with the study area. The variance η^2 of the Gaussian is set to be $d_T/4$ to cover approximately an area of 0.99 within the buffer radius of d_T around the store centre $\mathbf{l}_s$. Calculating the AUC for an arbitrary polygon as shown in Fig. 3.5, is computationally challenging. Henceforth I use the Monte Carlo method, where the samples are drawn from $\mathcal{N}(\mathbf{l}_s, \eta^2\mathbf{I})$ and reject them if outside the τ to calculate the fraction of samples.

The adjusted revenue $\widetilde{y_s}$ is formulated as the actual revenue weighted by the AUC:

$$\widetilde{y_s} = y_s \times \mathcal{A}. \tag{3.28}$$

This is applied for edge correction in the real-world data before fitting the BSIM.

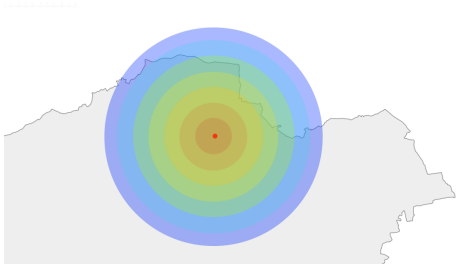


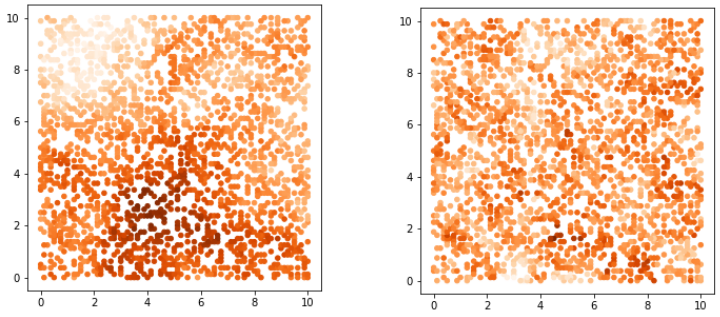
Figure 3.5: The red marker denotes a store at the edge of London. There may be customers who contributes to its revenue but not in the study area. Intersection of the radius and London map results in an arbitrary polygon.

3.3 Synthetic Experiments

A simulation study is designed to examine the inferences obtained from VI and MCMC methods under different synthetic settings characterised by an increasing number of stores and customers. I also compare the computational performance of the two methods by observing the run time of each fitted model. First, the data is simulated from a spatial process that closely matches the modeling framework introduced in Section 3.2, Eq. (3.2) with $\varepsilon_s = 0$. The process is defined as:

$$y_s|\boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma^2 \sim \mathcal{N}(r_s, \gamma^{-1}), \tag{3.29}$$

where the locations of stores and customers are simulated within a square. Two customer features are generated, one with a strong spatial correlation and the other with a moderate spatial correlation to closely reflect the real-world customer features as shown in Fig. 3.6. The store locations are randomly sampled within the same spatial boundaries used to sample the customers. Store features are sampled from a Gamma distribution ($\boldsymbol{\Phi} \sim \text{Gam}(1, 1)$) to represent features such as floorspace.

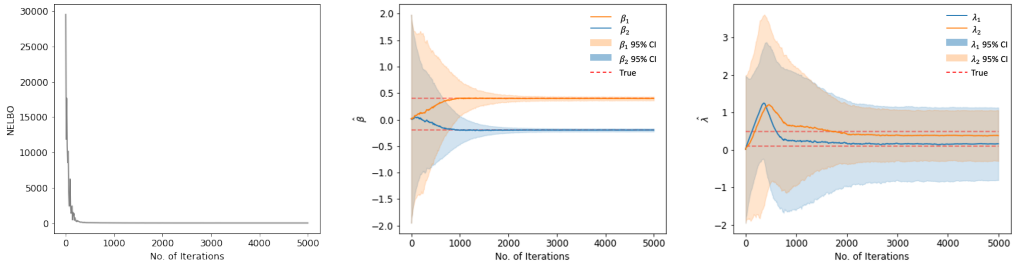


(a) (b)

Figure 3.6: Simulated Customer features for $N = 1000$ under two different spatial correlation structures to closely simulate the real-world scenarios: (a) Strong Spatial Correlation; (b) Moderate Spatial Correlation.

3.3.1 Parameter Estimation

For both VI and MCMC methods, all priors are chosen to be weakly informative to allow the data to drive the inference as illustrated in Table 3.3. VI optimisation algorithm is run over 5000 iterations where the convergence of the negative ELBO and parameters are illustrated in Fig. 3.7.



(a) (b) (c)

Figure 3.7: Convergence of the parameters over iterations in running the optimisation algorithm for VI : (a) negative ELBO, (b) β and (c) λ.

MCMC model is fitted using one chain with 5000 iterations by removing the first 2500 for warm-up, and every post-warm-up iteration is used for posterior samples. The diagnostics tests indicated that the chains have converged to a common distribution with R-hat close to one. The summary statistics for parameter estimates, and sampler diagnostics are presented in Table 3.1 and Fig. 3.8.

Table 3.1: MCMC summary statistics for parameter estimates, and sampler diagnostics

Parameter	Mean	SD	Quantile			n_eff †	R-hat
			2.5%	50%	97.5%		
β_1	-0.198	0.17	-0.235	-0.197	-0.166	982	1.000
β_2	0.400	0.021	0.358	0.399	0.447	1006	1.000
λ_1	0.185	0.547	-0.839	0.168	1.387	1201	1.001
λ_2	0.387	0.393	-0.313	0.361	1.269	1394	1.000
γ	1.908	0.304	0.562	1.759	4.054	1218	1.001

† provides a crude measure of effective sample size for each parameter.

The posterior distributions along with the prior distributions are visualised in Table 3.3 and parameter estimates are presented in Table 3.2. The results indicate that both methods approximate the posterior mean effectively and variational approximations of the posterior variance are lower than MCMC method.

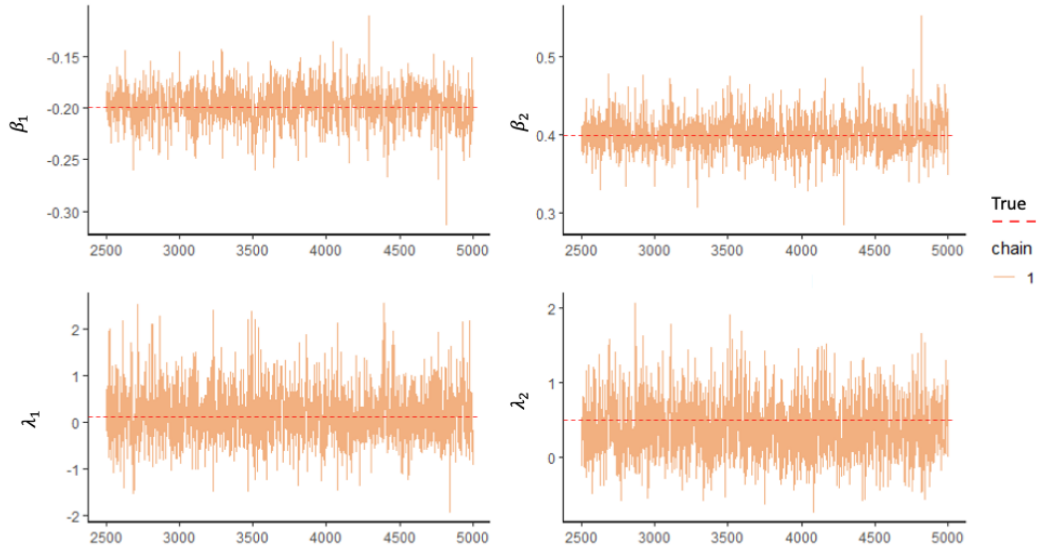


Figure 3.8: Traceplot corresponding to Markov chains, for a visual representation to inspect sampling behaviour and assess convergence.

Table 3.2: The first row indicates the True values of the parameters used to create the synthetic data, and the following rows display the first (Mean) and second moments (Standard deviation) along with its 95% quantile-based Credible Intervals (CI) for the posterior distributions for VI and MCMC methods.

		β_1	β_2	λ_1	λ_2	γ
True		-0.2	0.4	0.1	0.5	4
VI	Mean	-0.196	0.398	0.164	0.383	1.821
	Std	0.014	0.018	0.235	0.116	0.727
	CI	$(-0.224, -0.169)$	(0.362,0.434)	$(-0.296, 0.625)$	(0.156, 0.609)	(0.687, 3.499)
MCMC	Mean	-0.198	0.400	0.185	0.387	1.908
	Std	0.017	0.021	0.547	0.393	0.904
	CI	$(-0.235, -0.166)$	(0.358 , 0.447)	$(-0.839, 1.387)$	$(-0.313, 1.269)$	(0.562, 4.054)

The simulation process explained above is experimented under two different synthetic settings:

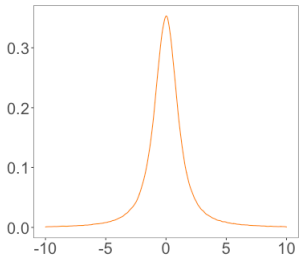
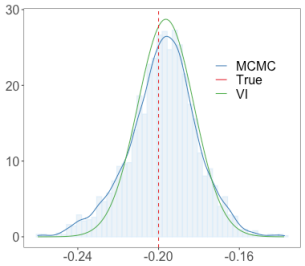
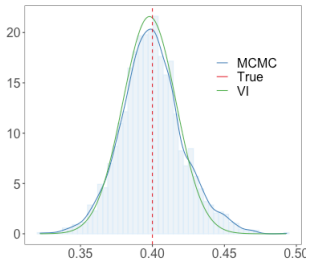
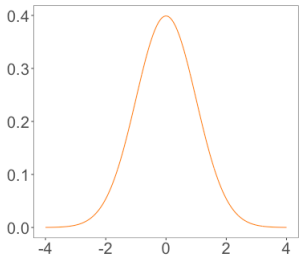
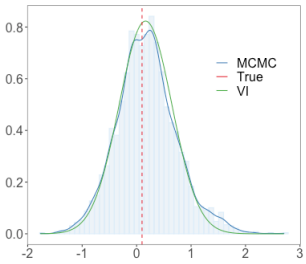
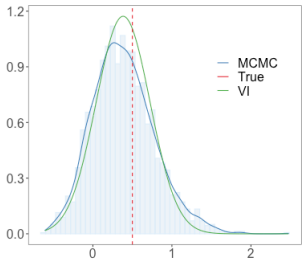
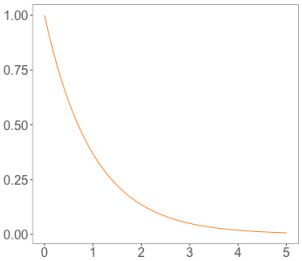
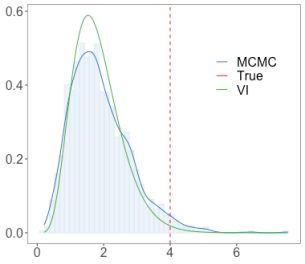
1. sim_1: 10 stores with 1000 individuals ($S = 10, N = 1000$)

2. sim_2: 50 stores with 2000 individuals ($S = 50, N = 2000$)

Random store locations are simulated to create 50 datasets and compare the performance across datasets using the posterior means of $\boldsymbol{\beta}, \lambda, \gamma$ and the 95% quantile-based credible intervals for each parameter from each fitted model. Three standard measures are used to compare the performance between MCMC and VI methods:

1. the bias, which measures the differences between the posterior mean

Table 3.3: Column one demonstrates the weakly informative prior distributions, and the following columns illustrate marginal posteriors of the interested parameters inferred by VI and MCMC. Synthetic experiment consists of 10 stores and 1000 customers ($S = 10$, $N = 1000$).

Prior	VI vs. MCMC	
$p(\boldsymbol{\beta}\|\alpha) \sim \mathcal{N}(0, \alpha^{-1}\mathbf{I})$ $p(\alpha) \sim \Gamma(1,1)$	β_1	β_2
		
$p(\lambda) \sim \mathcal{N}(0, \alpha^{-1}\mathbf{I})$	λ_1	λ_2
		
$p(\gamma) \sim \Gamma(1,1)$	γ	
		

from the model fit to dataset$_i$ ($\hat{\beta}_i$) and the true value of the parameter β, Bias $= \frac{1}{50}\sum_{i=1}^{50}(\hat{\beta}_i - \beta)$;

2. the mean-squared error (MSE), which takes the squared of the difference between posterior mean and true value, MSE $= \frac{1}{50}\sum_{i=1}^{50}(\hat{\beta}_i - \beta)^2$;

3. the coverage of the 95% quantile-based credible interval (CI) obtained from fitting the model to dataset$_i$, coverage $= \frac{1}{50}\sum_{i=1}^{50} I\left(\beta \in \text{credible interval}_i\right)$, where $I(\cdot)$ is the indicator function equal to 1 if the statement is true and 0 otherwise.

Table 3.4 and Table 3.5 show the results of the fitted models for the two synthetic settings, averaged across the 50 datasets. Both VI and MCMC

algorithms exhibit comparable performance in terms of bias, MSE, and coverage across both simulation studies. For sim_1, it is observed lower coverage for γ with the VI scheme. However, the coverage for γ is improved to one in the sim_2. Both λ and γ parameters result in a higher estimated MSE under both the simulation setting for VI and MCMC methods. This is an indication of the lack of identifiability in the parameters due to the flexibility in the model. The precision γ of the error term tends to be underestimated on average. Both models are fitted on an Intel Xeon CPU (3.5GHz and 32 GB of RAM). The run time of the VI algorithm is about five times faster than the MCMC algorithm in the simulation study. This is vital for our real-world data application, where the number of spatial locations is much larger than the synthetic settings. Overall the VI algorithm exhibited a reduced run time while providing good estimations and inference of the parameters of interest in this simulation study.

Table 3.4: VI and MCMC simulation study performance for $S = 10, N = 1000$.

Metric	Method	β_1	β_2	λ_1	λ_2	γ
Bias	VI	-0.002	0.004	0.258	0.110	-1.828
	MCMC	-0.002	0.004	0.265	0.116	-1.772
MSE	VI	0.000	0.000	0.130	0.051	3.467
	MCMC	0.000	0.42	0.130	0.049	3.276
Coverage	VI	0.94	0.96	1.	1.	0.44
	MCMC	0.96	0.98	1.	1.	0.94
Run time (s)		VI		MCMC		
		207		1064		

Table 3.5: VI and MCMC simulation study performance for $S = 50, N = 2000$.

Metric	Method	β_1	β_2	λ_1	λ_2	γ
Bias	VI	0.000	0.002	-0.338	0.352	-0.754
	MCMC	0.002	-0.001	-0.092	0.341	-0.734
MSE	VI	0.000	0.000	0.185	0.179	0.598
	MCMC	0.000	0.000	0.008	0.186	0.571
Coverage	VI	1.	0.94	0.84	0.94	1.
	MCMC	1.	1.	1.	0.857	1.
Run time (s)		VI		MCMC		
		1079		5280		

3.3.2 Model Predictions

The facilities operating in an area at time t and time $t + 1$ are simulated as illustrated in Fig. 3.9. At time $t + 1$ more stores are added and that leads to

more competition and loss of revenue for the existing facililities as compared in Fig. 3.9(c). First fitted the model for time t and estimated parameters are presented in Table 3.2.

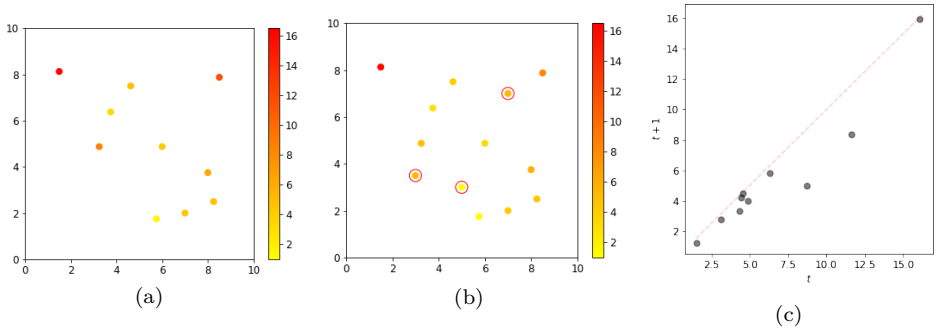


Figure 3.9: Synthetic setting of stores locations and their revenues indicated by the colour gradients at (a) time t and (b) time $t+1$, new stores are denoted by red colour circles. (c) Comparison of the revenues at time t and $t+1$.

Then model predictions are conducted for both time t and time $t+1$, and performance is evaluated with three standard metrics:

1. the Normalised Root-Mean-Squared Error (NRMSE), which measures the differences between the values predicted by a model ($\hat{\mathbf{Y}}$) and the values observed ($\mathbf{Y}$), $NRMSE) = \frac{\sqrt{E[\mathbf{Y}-\hat{\mathbf{Y}}]^2}}{E[\mathbf{Y}]}$;

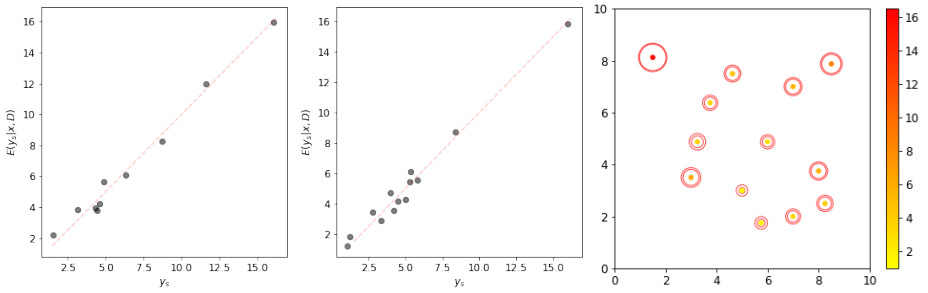
2. the R-squared, which is the ratio of the variance of the residuals (SS_{res}) and he variance of the observed $\mathbf{Y}$ (SS_{tot}), $\text{R}^2 = 1 - \frac{\text{SS}_{res}}{\text{SS}_{tot}}$.

3. the coverage of the 95% prediction intervals, that computes the actual coverage percentage of the prediction intervals on samples; larger the coverage, the better the model [80].

The model performance are summarised in Table 3.6 and visualised predictions against actual revenues in Fig. 3.10. Both time t and $t+1$ prediction performance are consistent with high R^2 and coverage of 100%. This indicates that the model can forecast the revenues of facilities with future changes in the spatial structure with good accuracy.

Table 3.6: R^2, NRMSE and prediction interval coverage for the BSIM .

Time	R^2	NRMSE	Coverage
t	0.98	0.08	100%
t+1	0.97	0.1	100%



(a) (b) (c)

Figure 3.10: Predicted revenue against actual revenue at: (a) time t and (b) time $t+1$. (c) The outer and inner rings show the 95% credible interval.

3.4 Model Comparison

In this section, the BSIM is compared against the prior literature in terms of the methodological advancement and compare the performance with a simulation study.

3.4.1 Methodological advancements

The spatial interaction models state that the perceived utility of a customer selecting a facility is positively related to the attraction and inversely related to distance, which is formulated as a ratio. As discussed in Chapter 2.2, the earliest developments formulated the relationship of distance as a power function [78] and later the exponential function is applied to represent the distance decay [152]. In the BSIM, a probability distribution is introduced to formulate the relationship between attraction and distance jointly. The variance of the distribution is modelled as a function of the facility's characteristics. Even though the Gaussian distribution is applied in this thesis, other probability distributions, such as the Beta distribution, can be used. This novel probabilistic approach provides more flexibility to capture the variation in utility perceived by customers with respect to the facilities.

Primarily the spatial interaction models are formulated to forecast the flows [5, 34, 135]. Even though it is common to find migration flows, acquiring transactional flows between customers is challenging. The BSIM overcomes this constraint by modelling aggregate level revenue or demand at stores. Henceforth, the BSIM can be applied to forecast the revenue at new sites, which is essential for businesses locating new facilities. Additionally, BSIM has the key advantage of providing interpretable inferences at the level of customers and stores. These inferences are beneficial for informed decision-making for property developers, planners, marketers, and business executives.

BSIM considerably improves existing classical spatial interaction models by formally addressing uncertainties arising in the modelling process via a

Bayesian framework. Recently, model parameters have been estimated using the known spatial structure in a Bayesian manner at the disaggregated level by assuming that the facilities had reached a stochastic equilibrium status [46]. Henceforth BSIM is a significant contribution to literature where it models aggregate level flows while providing probability densities of the revenue estimates at the store locations.

Traditionally the spatial interaction models are limited to experimenting with small datasets due to the scalability of the model calibration techniques that is addressed in this thesis. Classical models use entropy-maximising principles by applying numerical methods such as Newton-Raphson scheme [7, 34] or resorting to regression methods [8, 95, 105]. The recently developed Bayesian methods are limited to Markov Chain Monte Carlo MCMC [20, 46, 91] but does not scale up with large data. The parameters of the BSIM are estimated by applying the scalable variational inference technique. Hence this study stands out as the first scalable spatial interaction model with a Bayesian approach for estimating aggregate store-level revenue or demand.

3.4.2 Performance comparison

In order to demonstrate the performance of the BSIM, a comparison is conducted with one of the recently developed method known as the modified Huff model [95] using a simulation study. Various drawbacks of the traditional Huff type models are improved in the modified Huff model by considering both spatial competition and agglomeration concurrently. The purchase incidence is lower when multiple competitors exist, and in contrast, when there are multiple stores nearby, such as shopping malls, the purchase incidence tends to be higher, which is explained by the agglomeration effect [111]. This is graphically represented in Fig. 3.11. The BSIM does not explicitly account for the agglomeration effect, but the Gaussian distribution place at the centre of stores could learn their parameters to adjust for such consequences.

In the modified Huff model the probability a customer n visiting store s with size A_s is:

$$P_{ns} = \frac{A_s G_s^{\lambda} / d_{sn}^{\beta}}{\sum_s^S A_s G_s^{\lambda} / d_{sn}^{\beta}} \tag{3.30}$$

where G_s is the number of agglomeration stores within a radius of store s and its effect on shopping is reflected by parameter λ, d_{sn} is the distance between store s and customer n. The expected revenue at store s is:

$$E_s = \sum_n P_{ns} B_n C_s^{\gamma} \tag{3.31}$$

where B_n is the budgeted amount to spent, C_s is the number of competitors

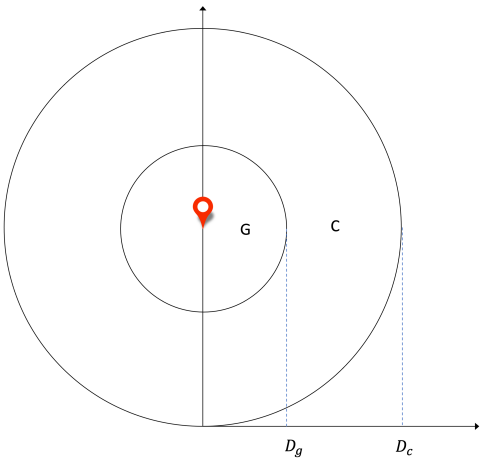


Figure 3.11: Agglomeration and competition areas in respect to the store shown in red pin. G is the agglomeration area and C is the service area (including G); D_g is the point that the agglomeration and competition forces are in balance; D_c is the radius of the service area.

within a radius of store s and γ reflects the effects of competition on shopping. A linear regression model is used to calibrate γ after applying Logarithms to Eq. (3.31). The remaining parameters (λ, β) are calibrated using an iterative optimisation approach by drawing values of the parameters incrementally from a range. Additionally the agglomeration and competition areas are evaluated against multiple radii. The final parameters are selected on the best fitted model based on R^2.

Two standard metrics are used to evaluate the performance: Table 3.7 displays the results for BSIM and Huff modified model. BSIM exhibits better performance across both the settings compared to the modified Huff model. An increase in NRMSE for both models are observed as the number of stores and customers increases. However, the R^2 remains unaffected and remains high, showing more robust performance for BSIM under both simulation settings compared to the modified Huff model.

Table 3.7: Performance of the simulation studies for BSIM and Huff modified model. $sim_1 : S = 10, N = 1000$ and $sim_2 : S = 50, N = 2000$.

		sim_1	sim_2
R^2	Model	0.98	0.94
	Modified Huff Model	0.77	0.30
NRMSE	Model	0.07	0.15
	Modified Huff Model	0.24	0.64

3.5 Summary

In this chapter, I have studied modelling the spatial interaction under uncertainty. First, a new approach is proposed to formulate the utility perceived by the customers through adopting a store-specific probability distribution. The distance and attraction attribute that influences the customer is represented with the probability distribution configurations. A Bayesian framework is proposed to model the spatial interactions that advance the existing model calibration methods in the literature. I have resorted to approximation techniques since the posterior distribution is analytically intractable.

A scalable variational inference framework is proposed that, while being significantly faster than competing Markov Chain Monte Carlo inference schemes, exhibits comparable performance in terms of parameters identification and uncertainty quantification. The benefits of BSIM is illustrated in various synthetic settings characterised by an increasing number of stores and customers. A detailed real-world application is demonstrated in Chapter 6.

In the next Chapter, the BSIM is extended to overcome the fixed demand assumption. The extended BSIM is used to identify the optimal facility location in the competitive location problem. A hierarchical search algorithm is proposed to solve the problem while presenting sampling techniques to overcome the requirement to identify an exhaustive set of potential locations.

Chapter 4

On the Competitive Facility Location problem with an extended Bayesian Spatial Interaction Model

4.1 Introduction

The geographical placement of a new business facility is of critical importance for commercial success. Growth in e-commerce continues to challenge the existence of physical retail stores. Therefore, it is essential to understand how customers interact with physical business facilities in order to design new commercial centres in competitive markets. I propose a modelling framework that accounts for customer behaviour to identify the optimal criteria for entering a new market or expanding its presence in a geographical region. This study aims to address three of the most pivotal questions facility planners face: the number of sites, their geographical locations, and design.

The formulation of optimal location models varies with the industry and purpose of the site. When locating facilities such as warehouses or manufacturing plants, the main focus is on the proximity to the customer, which is explained with proximity-based models [68]. In the context of locating emergency departments such as fire and ambulance services, the plan is to have the fewest number of sites so that all demand is covered within the stipulated maximum service response time, which is addressed with the location set covering problem [103, 140]. In contrast, *competitive facility location problems* emphasise industries such as retail businesses and commercial services, which consider competition when choosing their sites [15, 40]. These companies compete to attract customers buying power in a given area to capture market share.

As discussed in Chapter 2.2, one of the earliest probabilistic approaches for estimating market share was proposed by [78] based on the gravity model [126]. Huff's formulation states that the value or utility gained by a customer visiting a shopping centre is proportional to the store's floor space and inversely related to the power of the distance. Instead of the power function, it has been shown that exponential decay with additional store attraction better explains customer behaviour [39, 150]. Customers are assumed to patronise shopping centres based on their satisfaction indicated by a utility function [40]. The competitive location facility problem integrates the spatial interaction between customers and stores into the optimisation model according to utility models [10, 54].

In this study, the Bayesian spatial interaction model (BSIM), introduced in the previous chapter, is considered to integrate the spatial interactions in deciding the location. In BSIM, the utility gained by a customer visiting a facility is derived by evaluating the probability density function at the customers' location with respect to the underlying distribution centred on the store. Additionally, BSIM is based on a variational Bayesian approach, with key advantages of adaptability for large-scale problems and the ability to quantify aleatoric and epistemic uncertainty [84]. BSIM and in general spatial interaction models assume a fixed demand, but in most realistic situations, prices or availability of specific quality could affect the total number of customers patronising the stores or products. Hence I extend the BSIM to integrate such demand elasticities by adding dummy facilities as proposed by [89] and [41]. The extended BSIM method is adopted to model customer behaviour and estimate revenue generated at the new stores. My approach provides not only point estimates but also probability density estimates of revenues at optimal locations. Thus, the proposed competitive location modelling framework offers many advantages for decision-making over classical frequentist methods found in the literature.

In competitive facility location problems, the goal is to maximise the estimated market share or revenue of the business. Formulating the objective function of the optimisation model depends on the current state in the market of the company that searches for new sites. For instance, when a business with a chain of existing facilities plans to add several new stores, the objective is to increase market share captured by the chain, not just the additional site [42, 85]. I present the objective function of the optimisation problem considering three different scenarios: a company entering into a new market, a franchise expanding its presence in a competitive environment, and a business expanding in a monopolistic market. The objective function is maximised to choose the best locations and designs simultaneously from a set of potential sites and structures, in terms of store characteristics, within a certain budget.

In the process of establishing new facilities, the users are usually unable to provide an exhaustive set of potential sites, or this set is too large that it becomes computationally expensive. I propose a hierarchical search method that starts with a broad area and narrows the search to several regions to explore the neighbouring locations using a quadtree approach. The initial candidates are formed, ensuring that more potential sites are situated in areas with a high-density ratio between customer purchasing power and existing facilities. A non-parametric approach, kernel density estimation, is adopted to estimate the probability density functions [17]. According to the density ratio, the samples are generated from a multiresolution grid structure [130] and an inhomogeneous Poisson point process [93]. I evaluate the performance of these methods and regular grid sampling using synthetic experiments and demonstrate that the multiresolution grid structure outperforms other approaches.

My main contributions are: (a) propose a method to advance the BSIM in order to address one of the limitations by including lost demand in competitive environments to provide more realistic revenue estimates; (b) formulate an optimisation problem to simultaneously identify optimal facility locations and corresponding designs in competitive environments and provide probability density estimates of revenues at new sites; (c) propose a search algorithm based on the quadtree method to explore geographic regions of varying spatial resolution hierarchically.

4.2 Methodology

As discussed, the BSIM model assumes that the existing facilities capture the entire customer buying capacity, and there is no lost demand. This assumption is lifted by extending the BSIM in order to provide a more realistic formulation of customer choices and revenue predictions at business facilities. Next, I introduce the competitive facility location problem and a framework to search for optimal sites.

4.2.1 An extended Bayesian Spatial Interaction Model (BSIM)

Suppose there are N customers residing in location $\mathbf{m}_n \in \mathbb{R}^2$ having socio-demographic characteristics denoted by $\mathbf{v}_n \in \mathbb{R}^P$. Consider a set of available stores S where each store $s \in S$ located at $\mathbf{l}_s \in \mathbb{R}^2$ with store characteristics of $\boldsymbol{\phi}_s \in \mathbb{R}^D$. The customer $n \in N$ allocate their demand based on the utilities u_{ns} perceived by customer n for selecting each store $s \in S$. In the BSIM utilities are modelled by evaluating the probability density function (PDF) of truncated Gaussian distribution $f(d_{ns})$, centered on a facility $\boldsymbol{\mu}_s = \mathbf{l}_s$ and has a diagonal covariance matrix $\boldsymbol{\Sigma}_s = \sigma_s^2\mathbf{I}$ that indicates the store attraction.

This captures the likelihood for the n-th customer to visit the s-th store:

$$u_{ns} = f(d_{ns}) = \begin{cases} \dfrac{\exp\left(-d_{ns}^2/2\sigma_s^2\right)}{2\pi\sigma_s^2\left(1-\exp\left(-d_T^2/2\sigma_s^2\right)\right)}, & 0 \leq d_{ns} \leq d_T \\ 0, & otherwise. \end{cases} \tag{4.1}$$

where denotes d_{ns} the Euclidean distance between the customer and store locations $d_{ns} = ||\mathbf{m}_n - \mathbf{l}_s||_2$ and d_T is the maximum distance a customer would travel, beyond which the densities are set to zero in the truncated Gaussian distribution. Variance of the Gaussian σ_s^2 is formulated as a function of store characteristics $\boldsymbol{\phi}_s$ and the non-observable store characteristics $\varepsilon_s \in \mathbb{R}$;

$$\sigma_s^2 = \exp\left(\boldsymbol{\lambda}^\top \boldsymbol{\phi}_s + \varepsilon_s\right) \tag{4.2}$$

where $\boldsymbol{\lambda}$ represents a shared coefficients across the stores. Next the probability p_{ns}, for a customer n to visit a given store s is defined by,

$$p_{ns} = \frac{u_{ns}}{\sum_{j=1}^{S} u_{nj}} \tag{4.3}$$

BSIM and most spatial interaction models assume a fixed demand, but in most realistic situations, prices or availability of specific quality might affect the total number of customers using the facilities or products. I extend the BSIM to introduce elasticity of total demand as proposed by [89] and [41]. Henceforth, the model is advanced by introducing utility term u_{nd} assuming a dum'my facility in addition to the existing alternatives:

$$p_{ns} = \frac{u_{ns}}{\sum_{j=1}^{S} u_{nj} + u_{nd}} \tag{4.4}$$

It is now observed that the choice probabilities for a given customer (p_n) no longer always add up to unity.

$$p_n = \sum_{s=1}^{S} p_{ns} = \frac{\sum_{n=1}^{S} u_{ns}}{\sum_{j=1}^{S} u_{nj} + u_{nd}} \leq 1 \tag{4.5}$$

The dummy facility is assumed to be located at the same distance d_D for all customers. The distance d_D represents a reasonable extent ($d_D \leq d_T$) shoppers willing to travel. The revenue attracted by the dummy facility is considered to be the unsatisfied demand by the existing facilities. The variance of the Gaussian placed on the dummy facility is set to $\sigma_d^2 = d_T/4$, to obtain approximately 0.99 area under the curve within the maximum distance

a customer travel to a store. Hence the constant utility term u_{nd} is given by:

$$u_{nd} = \begin{cases} \dfrac{\exp\left(-d_D^2/2\sigma_d^2\right)}{2\pi\sigma_d^2\left(1-\exp\left(-d_T^2/2\sigma_d^2\right)\right)}, & 0 \leq d_D \leq d_T \\ 0, & otherwise. \end{cases} \tag{4.6}$$

The budgeted spending of a customer n is denoted by b_n is assumed to be a linear function of customer socio-demographics:

$$b_n = \boldsymbol{\beta}^\top \mathbf{v}_n \tag{4.7}$$

Finally the revenue or demand at a given store s is:

$$r_s = \sum_{n=1}^{N} b_n p_{ns} \tag{4.8}$$

Finally the complete data likelihood is:

$$p(\mathbf{Y}|\boldsymbol{\beta}, \lambda, \varepsilon, \gamma^{-1}) = \prod_{s=1}^{S} \mathcal{N}\left(y_s | \sum_{n=1}^{N} \boldsymbol{\beta}^\top \mathbf{v}_n \frac{u_{ns}}{\sum_{j=1}^{S} u_{nj} + u_{nd}}, \gamma^{-1}\right), \tag{4.9}$$

with $\mathbf{Y} = \{y_1, \ldots, y_S\}$ and the model assumes constant-variance (γ^{-1}) for the Gaussian noise. The posterior parameters of the extended BSIM are estimated using the variational inference approach similar to the Chapter 3.2. Customers are assumed to make their choices according to the extended BSIM, and the estimated parameters are used for the optimisation problem in locating new facilities.

4.2.2 Optimal facility location

Consider the problem where a company wants to find the optimal store facility to maximise the market share. An increase in revenue of new facilities is assumed to increase market share; thus, maximising revenue is equivalent to maximising market share. The optimisation problem aims to identify the optimal locations with store characteristics to gain maximum forecasted revenue within a set budget constraints. Consider an environment in which the customers are already served by existing stores L. Let $\tilde{L}$ denote the set of potential locations to open new facilities. For a given set of newly open stores $L^* \subseteq \tilde{L}$ the customer demand is split based on the utilities u_{nl} perceived by consumer n for selecting each facility $l \in L^*$. Suppose a discrete number of designs R are available and let $r \in 1 \ldots R$ represent a particular design. The features of a new store located at l with design r are denoted by $\boldsymbol{\phi}_{lr}$. Thus

the truncated Gaussian PDF denoted by $f(d_{nl_r})$ is:

$$f(d_{nl_r}) = \begin{cases} \dfrac{\exp\left(-d_{nl_r}^2/2\exp\left(\boldsymbol{\lambda}^\top \boldsymbol{\phi}_{lr}\right)\right)}{2\pi\sigma_s^2\left(1-\exp\left(-d_T^2/2\exp\left(\boldsymbol{\lambda}^\top \boldsymbol{\phi}_{lr}\right)\right)\right)}, & 0 \leq d_{nl} \leq d_T \\ 0, & otherwise. \end{cases} \tag{4.10}$$

where d_{nl_r} denotes the Euclidean distance between the customer and new store location l with design r.

Let x_{lr} be a binary variable set to one if and only if the company decides to locate a store at $l \in \tilde{L}$ with design r. Then the utility u_{nl} can be written as:

$$u_{nl} = \sum_{r=1}^{R} f(d_{nl_r}) x_{lr} \tag{4.11}$$

Consequently, the probability for customer n to visit new store l is calculated as:

$$p_{nl} = \frac{u_{nl}}{u_{nL} + u_{nd} + \sum_{l' \in \tilde{L}} u_{nl'}}, \tag{4.12}$$

where u_{nL} represents the total utility derived by customer n from all the existing stores. Total revenue generated by the new store locations L^* formulated by:

$$y_{L^*} = \sum_{l \in \tilde{L}} \sum_{n=1}^{N} b_n \frac{u_{nl}}{u_{nL} + u_{nd} + \sum_{l' \in \tilde{L}} u_{nl'}} \tag{4.13}$$

Let c_{lr} be the cost of locating a facility with design r at $l \in \tilde{L}$. Suppose the available budget for locating new facilities is $B \in \mathbb{R}$, and thus the budget constraint is obtained by:

$$\sum_{l \in \tilde{L}} \sum_{r=1}^{R} c_{lr} x_{lr} \leq B \tag{4.14}$$

Objective function

The objective function of the optimisation model depends on the current state in the market of the company that searches for new sites. Thus three unique objective functions are formulated and denoted by $\nu(x_{lr})$:

Case I: Consider a company that wants to find the optimal store location to enter a new market. The objective is to maximise the revenue of the new

facilities, and the objective function is expressed by:

$$\sum_{n=1}^{N} b_n \frac{\sum_{l\in\tilde{L}} \sum_{r=1}^{R} f(d_{nl_r}) x_{lr}}{u_{nL} + u_{nd} + \sum_{l\in\tilde{L}} \sum_{r=1}^{R} f(d_{nl_r}) x_{lr}} \tag{4.15}$$

Case II: Suppose a company already has a chain of existing facilities in a market $\hat{L}$, wants to build new stores to expand their presence. In this scenario, the company would wish to maximise the revenue of the new facility and make sure their existing facilities revenues are less affected. Henceforth the objective would be to maximise the total revenue of the current and new stores owned by the company. The objective function is:

$$\sum_{n=1}^{N} b_n \frac{\sum_{l'\in\hat{L}} u_{nl'} + \sum_{l\in\tilde{L}} \sum_{r=1}^{R} f(d_{nl_r}) x_{lr}}{u_{nL} + u_{nd} + \sum_{l\in\tilde{L}} \sum_{r=1}^{R} f(d_{nl_r}) x_{lr}} \tag{4.16}$$

Case III: The following scenario is where the market is a monopoly in which all the facilities are owned by one franchise. The objective would be to locate new facilities while optimising the total revenue generated from the market. The objective function is given by:

$$\sum_{n=1}^{N} b_n \frac{u_{nL} + \sum_{l\in\tilde{L}} \sum_{r=1}^{R} f(d_{nl_r}) x_{lr}}{u_{nL} + u_{nd} + \sum_{l\in\tilde{L}} \sum_{r=1}^{R} f(d_{nl_r}) x_{lr}} \tag{4.17}$$

Optimisation problem

Given the above definitions, I formulate the optimisation problem that is applicable for all three cases with the common constraints to find at most k number of locations to build new facilities within the given budget B to maximise the revenues:

$$\max_{x_{lr}} \quad \nu(x_{lr}) \tag{4.18}$$

$$\text{subject to:} \quad \sum_{l\in\tilde{L}} \sum_{r=1}^{R} c_{lr} x_{lr} \leq B \tag{4.19a}$$

$$\sum_{l\in\tilde{L}} \sum_{r=1}^{R} x_{lr} \leq k \tag{4.19b}$$

$$\sum_{r=1}^{R} x_{lr} \leq 1 \quad \text{for} \quad l \in \tilde{L} \tag{4.19c}$$

$$x_{lr} \in \{0,1\}, \quad \text{for} \quad r = 1 \ldots R; \quad l \in \tilde{L} \tag{4.19d}$$

where constraint (4.19a) is an upper limit for the total cost, (4.19b) limits

the maximum number of facilities and (4.19c) ensures multiple designs are not used for the same store. Since the objective function in all three cases is a sum of ratios with binary variables, the optimisation problem is identified as an integer nonlinear programming problem. The problem is related to the family of *multiple-choice knapsack problem* (MCKP) with generalised upper bound constraints, which is proven to be NP-hard [83]; hence our problem is NP-hard. The MCKP problem selects at most one item to pack into a knapsack from disjoint classes to maximise the sum of profits similar to our problem but differs by the objective function where it uses a sum of ratios. These types of problems are known as combinatorial optimisation problems, where the aim is to select a subset of the items to maximise the profit [153]. The optimisation problem is solved using constraint programming with the CP optimiser on IBM ILOG CPLEX studio 20.1.

4.2.3 Hierarchical search

In establishing a new facility, it is tedious for planners to provide an exhaustive set of candidate locations, or this set is so large that it is computationally expensive. A hierarchical search algorithm is proposed to start with potential locations from a broader region and narrow it down to explore neighbourhood locations. The algorithm executes a sequence of actions at several levels. The pseudo-code of the algorithm is presented in Algorithm 2.

Algorithm 2: Hierarchical search

load $\tilde{L}$; `// Load set of potential locations`
initialise $\mathcal{L}$; `// Create matrix` $\mathcal{L}$ `to save optimal locations` L^*
$\tau \leftarrow threshold$;
for *samples in* $\tilde{L}$ **do**
  $L^*, \nu \leftarrow$ findOptimalLocs($samples, B, k$) ;
  save L^* in $\mathcal{L}$;
end
$\Delta\nu \leftarrow \tau$;
while $\Delta\nu \geq \tau$ **do**
  $\nu_0 \leftarrow \nu$;
  $L^*, \nu \leftarrow$ findOptimalLocs($\mathcal{L}, B, k$) ;
  $\mathcal{L} \leftarrow$ getQuadtree(L^*) ;
  $\Delta\nu \leftarrow (\nu - \nu_0)/\nu_0$;
end

Three options are presented in the following sections to generate the initial set of candidate locations for the hierarchical search algorithm. Before executing the optimisation algorithm, the potential facility locations are split into random samples in the first level. Decomposing the larger matrix into smaller samples improves computational complexity in optimisation algorithms. Addi-

tionally, partitioning improves efficiency significantly in distributed computing environments. The solution at the first level contains optimal locations selected independently from each list. Subsequently, these optimal sites become the new potential locations for the next level. In addition to these sites, the neighbourhood locations are produced using the quadtree method, which is a tree data structure. The cells where the optimal locations were found are subdivided into four quadrants and use the midpoint as their neighbourhood locations. In the second level, search for the optimal locations and calculate its objective value. If the improvement of the objective value is larger than the given threshold, then the new optimal locations are recursively further decomposed and optimised with the new set of candidate locations until the improvement is smaller than the threshold.

Three sampling methods are proposed to generate the initial set of potential locations. The first method, the regular grid sampling approach, does not account for spatial variability. In contrast, the other two are density-based sampling methods; inhomogeneous Poisson point process and multiresolution accounts for spatial variability of customers and facilities.

Regular grid sampling

The regular grid sampling method does not account for the customer and facilities' spatial variability; thus, the candidate locations are uniformly distributed in space. The potential sites are generated using the midpoints of grid cells with a given resolution in a bounded region. The dimensions of the regular grids are a compromise between representation efficiency and computation overhead. A set of random samples are created to execute the hierarchical search parallelly using a stratified sampling approach. The data points are split into sub-regions or use statistical geographical boundaries and then sample from the subgroups independently.

Density-based sampling

I propose a density-based sampling method to create the initial candidate locations to overcome sampling errors in regular grid sampling. A non-parametric approach, kernel density estimation, is adopted for estimating the probability density function [17]. Given the existing facility locations $\mathbf{l}_s \in \mathbb{R}^2$, the density estimate at a point $x \in \mathbb{R}^2$ is given by:

$$f_s(x) = \frac{1}{Sh} \sum_{s=1}^{S} K\left(\frac{x - \mathbf{l}_s}{h}\right) \tag{4.20}$$

where $K(\cdot)$ is a kernel function, chosen to be a Gaussian kernel with bandwidth parameter h, optimally selected according to [133]. The spending power

b_n (Eq. 4.7) is unevenly distributed at customer locations $\mathbf{m}_n$. Hence a weighted kernel density estimator is considered to model customer spending capacity [60]. The spending capacity b_n at each customer location $\mathbf{m}_n$ is normalised and denoted by w_n, so they add up to one. The weighed density estimate function is given by:

$$f_n(x) = \frac{1}{Nh} \sum_{n=1}^{N} w_n K \left(\frac{x - \mathbf{m}_n}{h} \right) \tag{4.21}$$

A ratio $f_r(x)$ between the density estimates is calculated. This provides an indicator of how dense the area is in terms of customers spending power compared to the available facilities:

$$f_r(x) = \frac{f_n(x)}{f_s(x)} \tag{4.22}$$

Two sampling methods are proposed using the estimated density ratio.

Sampling with inhomogeneous Poisson point process
The potential set of locations is simulated using the inhomogeneous Poisson points process (IPPP) to have many locations in the sample from regions with high intensity of the ratio $f_r(x)$. In a homogeneous Poisson process with intensity λ, the number of events η in any bounded region A is Poisson distributed with mean $\lambda|A|$ where $|A|$ denotes the area of A [28]. In contrast, the intensity function of an inhomogeneous Poisson process is a nonconstant function $\lambda(x)$ of spatial location $x \in \mathbb{R}^2$.

IPPP is simulated through [93] thinning algorithm. First a random number η^* is obtained from a Poisson distribution with mean $\mu(A) = \int_A \lambda(x)dx$. Next, a homogeneous Poisson point process is simulated with intensity value λ^* which is an upper bound of the intensity function $\lambda(x)$. For this the maximum of the ratio between the density estimates, $\lambda^* = \max f_r(x)$ is used. Finally, points x^* of the homogeneous process is thinned according to $f_r(x^*)/\lambda^*$ (i.e. each point x^* is deleted independently if a uniform(0,1) random number is greater than $f_r(x*)/\lambda^*$) which results in a IPPP forming the candidate lo cations for the hierarchical search. The second level of the hierarchical search does not continue recursively since the grids are not used to generate data, unlike the other two proposed methods.

Sampling with Multiresolution grid structure
The multiresolution depth grid is created in the proposed approach based on the estimated density ratio $f_r(x)$. First, $f_r(x)$ is estimated on a fine meshgrid created in the study region. Next, create a regular grid and calculate the

average μ_r, of $f_r(x)$ within each cell. Compute the q-quantiles of the μ_r, and assign to which quantile each cell resides. This represents the number of iterations to decompose each cell into four smaller sub-blocks. The midpoint of sub-blocks is used as the candidate locations. The dimension of the regular grid and depth of resolution (q) is a compromise between representation efficiency and computation overhead. The pseudo-code of the method is presented in Algorithm 3.

Algorithm 3: Multiresoltion grid structure

```
x ← constructPointsMeshgrid(region);
grid ← decompose(region, m) ;        // decompose region into m
 sub-blocks
foreach c_i in grid do                  // for each cell c_i in the grid
    μ_r ← mean(f_r(x)) ;     // Mean of f_r(x) of points in cell c_i
    save μ_r in c_i;
end
q̂ ← max(μ_r)/q ;              // q denotes the depth of resolution
foreach c_i in grid do
    points ← midpoint(c_i) ;                  // Midpoints of cell c_i
    save points in out ;
    for j = 1 to q do
        if (j − 1) × q̂ ≥ μ_r ≤ j × q̂ then
            for g = 1 to j do
                c_i ←decompose(c_i, 4) ;        // Decompose c_i into 4
                 square submatrices and repeat recursively
                points ← midpoint(c_i) ;
                save points in out ;
            end
            exit
        end
    end
end
```

4.3 Synthetic Experiments

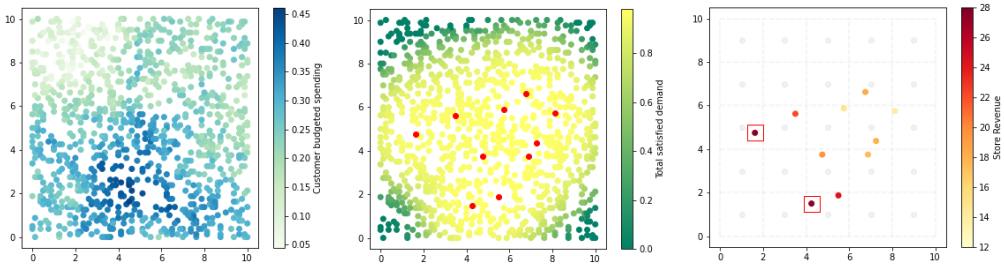
A simulation study is designed to experiment with the optimisation problem using the three objective functions introduced and compare the performance using the three sampling methods proposed in Section 4.2. The computational performance of the methods is compared by observing each optimisation problem's run time. All the experiments are executed on an Intel Core i5 CPU (2.3 GHz Dual-Core and 8 GB of RAM).

First, data is simulated from a spatial process that closely matches the extended BSIM framework introduced in Section 4.2 with the dummy facilities.

The process is defined as:

$$y_s|\boldsymbol{\beta}, \boldsymbol{\lambda}, \sigma^2 \sim \mathcal{N}(r_s, \gamma^{-1}), \tag{4.23}$$

where a reasonable distance that a customer is willing to travel is assumed to be half of the maximum extent prepared to travel ($d_D = d_T/2$). The locations of stores and customers are simulated within a square. Customer budgeted spending is generated, with a strong spatial correlation where rich and poor areas are demonstrated to reflect the real-world scenario closely, as shown in Fig. 4.1(a). The customers' satisfied demand (p_n) from the existing stores are shown in Fig. 4.1(b). The store locations are randomly sampled, and their current revenue is displayed in Fig. 4.1(c). Two possible designs ($r = 2$), say small and large facility structures, are assumed to be available for development. Suppose the cost of a large building is six times the smaller facility, and the cost of each design (c_{lr}) remains unchanged despite the locations.



(a) (b) (c)

Figure 4.1: (a) Simulated customer locations (N = 1000) and budgeted spending (colour gradient); (b) Satisfied customer demand (colour gradient) and the existing stores (red) (c) Revenue of the existing stores (colour gradient) and potential store locations (grey).

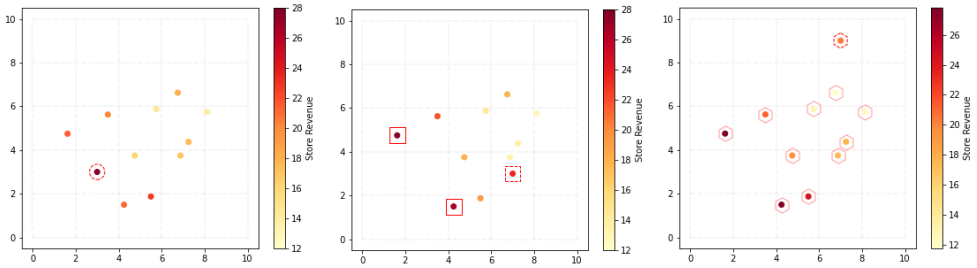
4.3.1 Demonstration of the optimal facility location with varying objective functions

Given the above synthetic setting, the optimisation problem is solved to find the optimal location for one new facility with a budget of ten($B = 10$) for the three objective functions discussed in Section 4.2. A regular grid sampling approach is used to generate the potential facility locations, as presented in Fig. 4.1(c). The results of the optimisation problem with the objective function in case I (Eq. 4.15), a company entering the market for the first time, the new facility is to be located in the area with the wealthiest customers generating the highest revenue compared to all the facilities (Fig. 4.2(a)). In case II (Eq. 4.16), a franchise opening a new facility, the optimal location moves away from the other facilities in the chain as displayed in Fig. 4.2(b). Revenue of the new facility reduces compared to case I, but the total sales of the chain facilities

Table 4.1: Revenue of the existing and optimal facilities

		Optimisation		
	Existing	Case I	Case II	Case III
Total revenue of all stores	204.3	210.8	209.9	**213.6**
Total revenue of chain stores	55.4	70.3	**77.7**	75.9
Revenue of new store		**27.6**	23.5	20.4

are increased, as shown in Table. 4.1. Finally, in case III (Eq. 4.17), when opening a new facility in a monopolistic market, the new store locates away from all the existing facilities (Fig. 4.2(c)) to gain additional sales to maximise the total revenue of all the facilities. Total revenue shows the highest while the sales at the new facility show the lowest compared to other cases (Table. 4.1). In all three cases, the optimal facility design is large size.



(a) (b) (c)

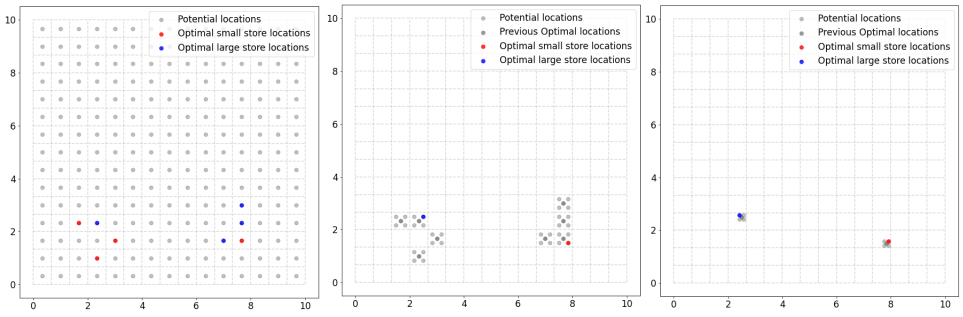
Figure 4.2: The experiment is to find the optimal location for a new facility under three different objectives: (a) Maximise the revenue of the new facility (Eq. 4.15). (b) Maximise the revenue of all facilities owned by the franchise (Eq. 4.16). Square indicates the existing facilities owned by the franchise. (c) Maximise the revenue of all facilities in the market (Eq. 4.17). Hexagons indicate that all facilities owned by the same company. The optimal location is shown within the red colour dashed circle, square and hexagon.

4.3.2 Evaluation of sampling methods for the hierarchical search

I experiment with the hierarchical search using the three sampling methods proposed in Section 4.2. The synthetic setting remains as described above. Experiments are performed with the objective function where a new company is entering the market (Eq. 4.15) and looking to establish two facilities with a budget of 10. The threshold of the hierarchical search for the recursive stage is set to 0.01, meaning the objective function should increase by more than 1% to continue the search.

Regular grid sampling

A regular grid with dimensions of 10×10 is used to generate the initial candidate locations. The locations are split into four samples using the stratified sampling method. The optimisation problem is solved independently for each sample to identify two optimal sites forming eight in total, as shown in Fig. 4.3(a). For the next level of the hierarchical search, the neighbourhood locations are produced using the quadtree method forming 40 potential sites as reported in Fig. 4.3(b). The recursive algorithm stops after two iterations producing two locations to establish the new facilities with the two designs.

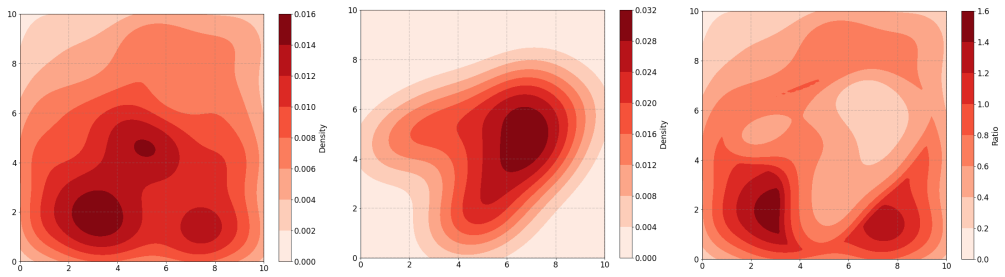


(a) (b) (c)

Figure 4.3: Visual progression for regular grid sampling for hierarchical search. (a) Initial candidate locations generated from 10×10 regular grid. Eight optimal locations are found from each sample producing one small and large design facilities. (b) Neighbourhood locations for the optimal locations generated from the previous step and the new optimal locations. (c) Final optimal locations are derived from ten potential locations.

Density-based sampling

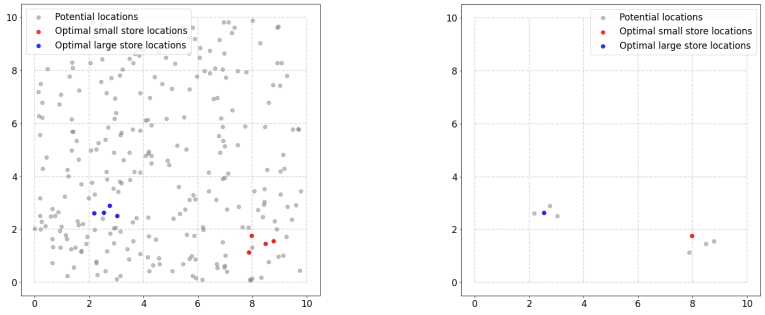
The initial step for the density-based sampling method is to fit the kernel density functions for customer spending budget and the store locations. Fig. 4.4 demonstrates the density contour plot generated for customer purchasing power, store locations and the ratio of the two density estimates in the area. A 100×100 meshgrid is used to estimate the density and calculate the ratio.



(a) (b) (c)

Figure 4.4: Demonstrates the density contour plot generated for (a) customer spending, (b) store locations (c) ratio of the two density estimates in the area.

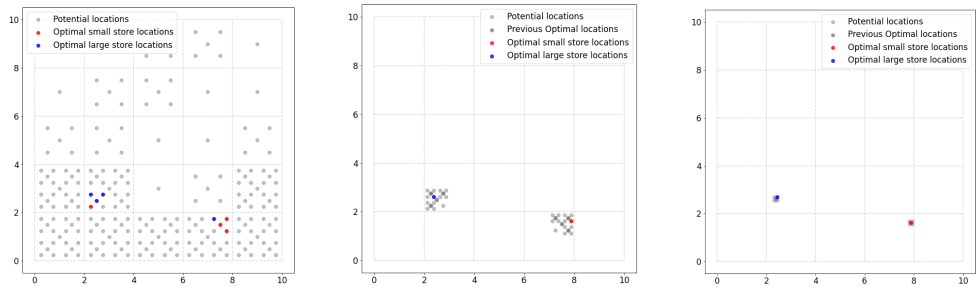
Sampling with inhomogeneous Poisson point process: The maximum estimated ratio from the meshgrid is considered as the λ^* for the IPPP. Four random samples are generated from IPPP and solved the optimisation problem independently to identify eight optimal facilities as displayed in Fig. 4.5. These eight locations become the potential sites for the final iteration to find the optimal facilities.



(a) (b)

Figure 4.5: Visual progression for IPPP sampling for hierarchical search.(a) Random samples are generated from the IPPP as the potential locations for the optimisation problem. Eight optimal facilities are found, with each sample producing one small and large facility location. (b) The optimal locations of the previous stage becomes the candidate sites for the second level from which the optimal locations are identified.

Multiresolution sampling: A regular grid of 5×5 is created and calculate the average within each cell using the estimated density ratios from the meshgrid. The resolution depth is chosen to be three and calculate 3-quantiles of the average values to decide the resolution of each cell. Fig. 4.6 (a) presents the multiresolution samples used as the potential locations. The set of candidates is split into four random samples and solve the optimisation problem independently. The algorithm stops after two iterations providing the optimal facilities, as shown in Fig. 4.6 (c).



(a) (b) (c)

Figure 4.6: The progression of the multiresolution sampling method to find the optimal locations. (a) Multiresolution samples for 5×5 grids with a depth of three. (b) Neighbourhood locations for the optimal sites generated from the previous step and the new optimal locations. (c) Final optimal locations are derived from ten potential sites.

Comparison between sampling methods

The results are compared in terms of the final objective values and the run time of the hierarchical search for the setup described above. All three methods show consistent results, whereas the multiresolution approach shows marginally higher optimal revenue as reported in Table. 4.2. The optimal locations for all three methods are in the same regions, whereas the large store is located on the left and the small store on the right side.

Table 4.2: Results of the hierarchical search with the sampling methods

Sampling Method	Starting number of locations	Objective Value	Run time (s)
Regular grid	225 (grids = 15x15)	53.83	231
IPPP	271 (samples = 4)	53.78	240
Multiresolution	217 (depth = 3)	**53.85**	275

The comparison is extended by simulating the experiment 100 times. Datasets are created by generating random store locations while keeping the same setup for the customers described in the simulation study. The initial number of candidate locations for the three sampling methods are experimented with at two levels to demonstrate the behaviour based on the initial sample size. The performance are compared across three sampling methods and the initial sample size. The Table. 4.3 presents a summary of the results.

Table 4.3: Performance comparison between sample methods

Sampling Method	Starting number of locations	Number of times with best Objective value	Average objective value	Average run time
Regular grid	64 (grids = 8x8)	41	58.23	33
	225 (grids = 15x15)	47	58.30	252
IPPP	76 (samples = 1)	1	57.23	57
	262 (samples = 4)	0	58.02	288
Multiresolution	73 (depth = 2)	**58**	58.25	38
	217 (depth = 3)	53	**58.31**	241

The average of the objective value for each sampling method is marginally improved as the starting number of candidate locations for the hierarchical search increases. The multiresolution sampling method has obtained the highest objective value 58 times, while the regular grid method has obtained this 41

times for the setting with low number of starting locations. However, with the increase in the starting number of candidate locations, the comparison between the regular grid and multiresolution methods are narrowed. This implies that increasing the regular grid method's resolution could achieve as good a result as the multiresolution method. IPPP has performed comparatively poorly with only ones obtaining the best objective function. This could be because the IPPP approach does not recursively evaluate neighbouring locations. Since IPPP run time is higher than the other methods, an approach is not considered to explore the neighbouring sites. It can be concluded that the multiresolution sampling method could produce better results with a low number of starting locations while being efficient.

4.4 Summary

In this chapter, I have studied the competitive facility location problem that typically arises when businesses plan to enter a new market or expand their presence in an environment with existing competitors. A mathematical modelling framework is formulated to simultaneously identify the location and design of new stores in order to maximise revenue in a given geographical region. In doing so, the Bayesian spatial interaction model (BSIM) is extended by incorporating demand elasticity, thus providing more realistic revenue estimates. Solving the underlying allocation optimisation problem requires the provision of an exhaustive set of potential sites, which is difficult in practice. Instead, a search algorithm is introduced based on the quadtree method to overcome this challenge by hierarchically exploring geographic regions of varying spatial resolution. Different sampling techniques are proposed to generate the initial set of candidate locations for the algorithm: regular and multiresolution grid structures and inhomogeneous Poisson point processes. The multiresolution approach based on kernel density estimates is proven to be the most competitive performance.

The sampling techniques do not provide an exhaustive list or may not provide the true optimal sites that could build the facilities. But these optimal sites indicate the areas that need to be explored with real opportunities based on different criteria such as availability and size of land. The introduced framework could be valuable for decision-makers in companies, property developers, planners etc.

The framework is put into practice with two real-world case studies in Chapter 6. The challenges and limitations in applying to real-world problems are discussed along with its results. The next chapter introduces real-world, large-scale datasets required for real-world experiments.

Chapter 5

Introducing large scale geo-spatial datasets

5.1 Introduction

In the literature, experiments on spatial interaction modelling are limited to small synthetic datasets or real-world aggregated data since acquiring granular level real-world data is usually expensive [2, 14]. This is also common in competitive facility location literature, where the applications are limited to identifying optimal facility locations within a synthetic setting or small geographic regions [10, 41, 54]. To address these constraints, I create a dataset that includes variables observed at a granular level for public houses (pubs), supermarkets and customer zones. This is performed by combining large geospatial and non-geospatial data from open and commercial data sources. Additionally, I gather customer reviews from Google's customer rating API, which covers a broader audience compared to the traditional survey methods found in the literature [39].

My main contributions are: (a) constructed an unprecedented real-world large spatial dataset for over 1500 Pubs in Greater London to demonstrate revenue, physical store features, surrounding characteristics and customer ratings. (b) introduce a dataset with approximated revenues and store capacity for the nine largest supermarket chains in the UK; (c) over 150,000 postcodes, most granular administrative level, data set is compiled for Greater London to represent customer zones and characteristics.

5.2 Compilation of pubs and supermarkets datasets

Accessing revenue generated at the business locations are nearly impossible due to the confidential and competitive nature of the data. Hence in this thesis, I explore two industries, pubs and supermarkets, that have a close relationship

between revenues and rateable values published by Valuation Office Agency (VOA) [143]. This section presents a comprehensive description of the data and the steps to compile the two new datasets.

5.2.1 Non-Domestic properties

The Valuation Office Agency (VOA) [143] maintains rateable values, also known as business rates, of around 2 million non-domestic properties in England and Wales. The latest rating list was compiled in April 2017, and the next publication is due in 2022. The rateable value of business properties is usually adjusted every five years to reflect changes in the property market. The most common valuation method is the open market annual rental value of the property. Each local billing authority is responsible for compiling and maintaining the local rating list. The regular site and building survey support the majority of the properties rateable value. The local councils multiply the rateable value with the multiplier set by the VOA to calculate the business rates of non-domestic properties. The complete list can be downloaded in the CSV format from the VOA website [143] and are protected by Crown Copyright and Crown Database rights. Variables that are provided in the dataset is presented in Table 5.1.

The non-domestic properties include properties or land that are not solely used for residential. Each property is classified into over 300 categories (schools, pubs, hotels, food stores, nightclubs etc.). 80% of the non-domestic properties in England are represented by only 4% of the categories. The frequency distribution of these 15 categories is shown in the figure 5.1.

5.2.2 Properties geo-coordinates

The Ordnance Survey Addressbase premium [113] is the most comprehensive address data set for the UK, containing approximately 40 million addresses. Each property has a Unique Property Reference Number (UPRN) and is classified as either commercial or residential, and further classified into over 500 categories. The data set provides the spatial point coordinates for each property. This is a commercial proprietary product from the Ordnance survey. The spatial database is queried using PostGIS.

5.2.3 Properties boundary Polygons

This dataset provides the indicative shape and position of each boundary of a registered title for land and property in England and Wales [72]. Each title is either freehold or a leasehold with a unique title number with at least one index polygon. There are more than 25 million titles and 28 million polygons. The area of the title polygon gives the size of the land. The dataset is published

Table 5.1: Variables in VOA data

	Variable	Type	Description
1	Billing Authority Code	Character	Code representing Billing Authority
2	Primary And Secondary Description Code	Character	Code providing a high level description of the property
3	Unique Address Reference Number UARN	Number	VOA Internal key used to link information about the same hereditament
4	Full Property Identifier	Character	Location of property as shown in Rating List (Usually less than 100 characters).
5	Firms Name	Character	The name of the company
6	Number Or Name	Number	Number and or name of the hereditament
7	Street	Character	Name of Street
8	Town	Character	Name of Town
9	Postal District	Character	The Postal District
10	County	Character	Name of County
11	Postcode	Character	Postcode of hereditament as recorded by VOA
12	Effective Date	Date	Format DD-MON-YYYY Date the current assessment came into effect.
13	Rateable Value	Number	Rateable value is an assessment of the open market rental value of the property on the prescribed valuation date.
14	List Alteration Date	Date	Format DD-MON-YYYY. Date on which List entry was created or last amended
15	SCAT Code And Suffix	Character	Code used by the VOA to group properties together for operational purposes

as a commercial dataset by HM Land Registry. Variables that are provided in the dataset is presented in Table 5.2.

5.2.4 Topographic data

The Ordnance Survey MasterMap Topography Layer provides access to the most detailed, current, and comprehensive dataset of Great Britain's landscape [114]. Each record in the database offers geometric position and shape on Earth and its related attributes. This helps to identify the landscape, building footprint, and heights. Ordnance Survey develops, manage, and maintain the data within one of the world's largest spatial databases [114]. Real-world topographic features are represented by points, lines, and polygons, along with a unique identifier. The dataset size is over 40GB and is in the Geography

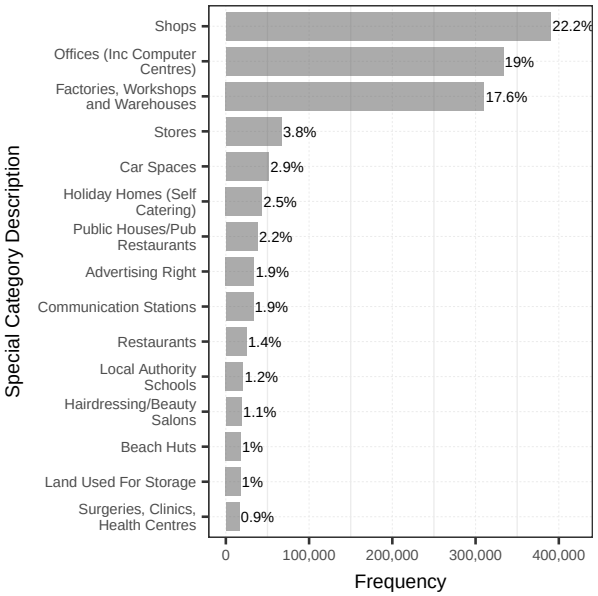


Figure 5.1: Frequency distribution for categroies that account for 80% of non-domestic properties in England.

Mark-up Language (GML) format.

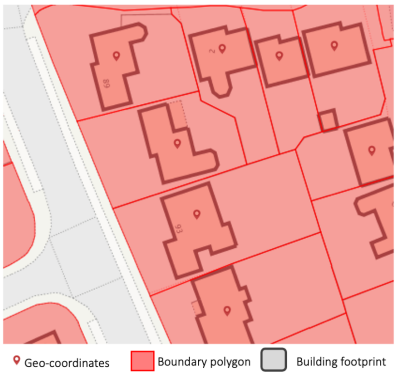


Figure 5.2: Illustration of proporties geo-coordinates, boudary polygon and topographics

5.2.5 Property Ownership

The ownership dataset published by HM Land Registry provides details of the registered properties in England and Wales that UK companies own [73]. The title number is directly linked to the national polygon dataset and is published monthly that is accessible for free in CSV format. The variables that are provided in the dataset is presented in Table 5.3.

Table 5.2: Variables in National polygon data

	Variable	Type	Description
1	Shape	Geometry	Geometry of the Index Polygon must be a single area
2	Poly_ID	Number	Unique polygon reference.
3	Title_No	Character	Unique number which identifies a registered title to land
4	Insert	Date	Date on which the polygon in the title was initially created on the index map
5	Update	Date	Date on which all or part of the title was last updated
6	Vers_No	Number	Version of Poly_ID
7	Rec_Status	Character	Identifier to describe status of the polygon. Added (A), Changed (C), Deleted (D)

5.2.6 London Business rates

In this thesis, the real-world applications are concerned with Greater London, England's capital and largest city. The non-domestic properties (5.2.1) and address geo-coordinate datasets 5.2.2 are joined using the cross-reference to develop a comprehensive spatial point dataset with rateable values. The spatial intersect between the property data set and statistical spatial boundary for London is used to subset the data set to filter London's non-domestic properties. Business rates are charged from 277,906 non-domestic properties in London. 90% of the non-domestic properties in London are represented by 16 of the categories, and the frequency distribution is presented in figure 5.3.

The rateable value ranges between £41 and £212.4 million, with an average value of £63,461. The log transformation is used on rateable values to overcome the skewness of the data. Figure 5.4. shows the variability in distributions for the log of rateable value for each of the 16 major categories.

5.2.7 External store characteristics

In addition to the internal store characteristics, the external features are extracted to use in the model. These features are concerned with explaining the urban environment of the stores are located. External characteristics are explained by denoting if it is placed in a major town, the closest distance to public transport access points [35], tourist attractions [71] and sports facilities.

Access points to public transport

The National Public Transport Access Nodes (NaPTAN) database lists all points of access to public transport in Great Britain. It records approximately

Table 5.3: Variables in the Corporate ownership dataset

	Variable	Type	Description
1	Title number	Character	Unique number which identifies a registered title to land or a caution against first registration.
2	Tenure	Character	Freehold or leasehold.
3	Property Address	Character	The unformatted address in the register.
4	District	Character	Name of an administrative district.
5	County	Character	Name of current county in England and Wales.
6	Region	Character	Name of a geographic region.
7	Postcode	Character	It is part of a coding system created and used by the post office across the UK.
8	Price Paid	Character	The sale price stated on the transfer deed
9	Proprietor name	Character	Non-private Individual Name. Given upto 3 owners of the title.
10	Company Registration	Character	A unique identifier assigned to a company when it is registered at Companies' House.
11	Proprietorship Category	Character	Text which describes the category in which a name falls.
12	Proprietor address	Character	Register address string.
13	Date Proprietor added	Date	The date a proprietor was added to the register.
14	Additional Proprietor Indicator	Character	Indicates if there are other proprietors in the register.

400,000 bus stops across England, Scotland and Wales, and other transport terminals, including rail stations and airports [35]. The dataset can be accessed openly and downloaded as a CSV. The given geo-coordinates are then converted into a spatial database and stored in PostgreSQL. This dataset is subject to the Open Government Licence.

Places of Interest

The tourist attractions such as monuments, world heritage sites, and parks and gardens spatial data are available on 'Historic England' under listing datasets [71]. The files are in the format of shapefiles and provide the locations as polygons. This dataset is subject to the Open Government Licence. Additionally, the locations of sports facilities are taken from the OS Addressbase data (5.2.2).

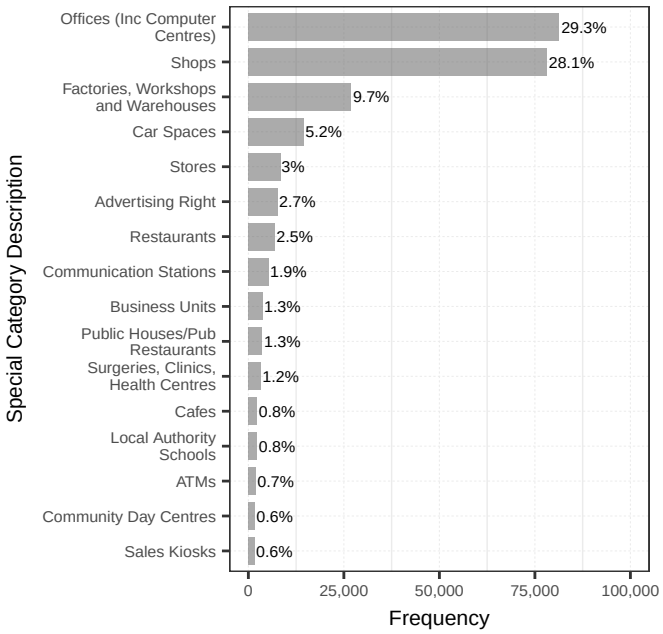


Figure 5.3: Frequency distribution of the categories that represent 90% of Non-domestic properties in London.

English town centres

The English town centres are provided as polygons in shapefile format [101]. The properties are spatially joined to obtain if it is placed within a town. The output is generated as a binary variable. The spatial distribution of the English towns is presented in Fig. 5.7.

5.2.8 Customer ratings

The customer rating is an important aspect to demonstrate store attractiveness. I have strengthened the store characteristics by including the customer reviews on Google [63]. People can write reviews and rate the places voluntarily on Google maps. The ratings are then aggregated and shown to the public. Obtaining data from a wider audience adds significant value compared to traditional survey methods. The Google Places API provides access to this data at a cost. The process applied in compiling the dataset is presented in Fig. 5.8.

5.2.9 Dataset with characteristics of pubs in Greater London

The calculation of the rateable values of pubs is different from other categories. In contrast, the rateable value of pubs is based on the annual level of trade (excluding VAT) that a pub is expected to gain if operated in a reasonably efficient way [142]. Hence the rateable value is a good proxy of the pub revenues. There are 40,000 pubs recorded in non-domestic properties dataset for

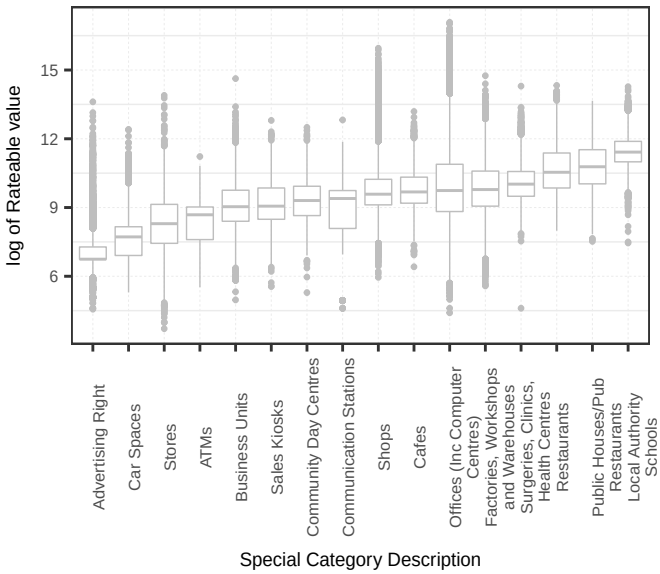


Figure 5.4: Frequency distribution of Rateable values in London.

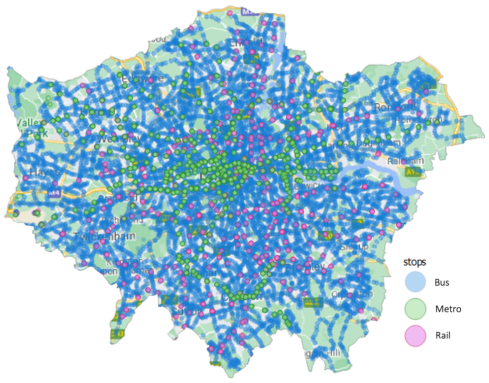


Figure 5.5: Illustration of public transportation access points in London.

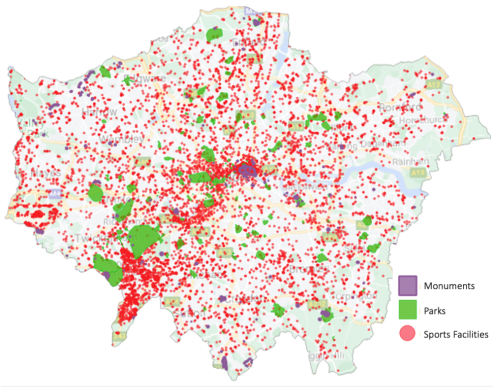


Figure 5.6: Illustration of the places of interest in Greater London.

England and Wales, and 3,534 in Greater London. The spatial distribution of

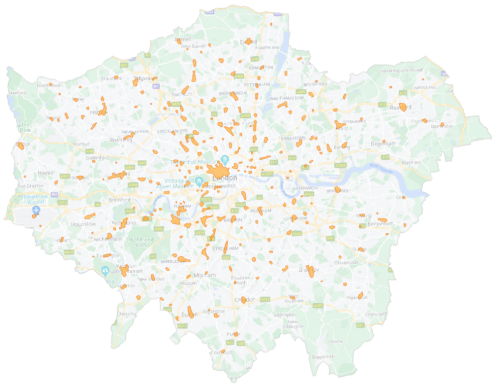


Figure 5.7: English town centres in Greater London.

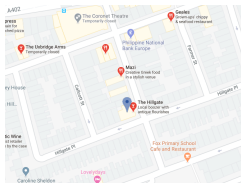
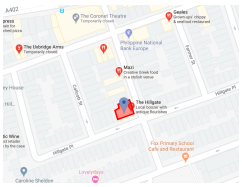
Data processing – Google Place data			Variables
Geo-locations for the pubs are taken from the dataset created from the initial data processing step.	Using the Google Place API, near by search option the locations with keyword 'pub' list is read in the json format. This is then transformed to a spatial table using the geo-coordinates.	The correct google pin is identified by spatial join with the title polygon of the respective pub.	• Name (The place's name) • Latitude • Longitude • Place ID • Vicinity (A simplified address for the place) • Price level • Rating (from 0.0 to 5.0, based on aggregated user reviews) • Types (e.g.: bar, restaurant, food) • Total number of ratings
			

Figure 5.8: Diagram demonstrate the steps of extracting Google ratings for Pubs. Since there is no direct cross-reference between Google and other datasets, I have employed spatial joining to link data.

pubs across England is shown in Fig. 5.9.

A dataset around pubs is compiled to demonstrate internal and external characteristics for each facility by combining the datasets described early in the chapter. This is accomplished by first spatially joining the polygon of the land [72] with locations of stores and next spatially joining the polygon of the footprint from Mastermaps. I have strengthened the store attractiveness measures by using the customer reviews on Google [63]. The flow diagram in Fig. 5.10 presents the process built to extract the store features. The summary statistics of the store features compiled are presented in Table 5.4.

5.2.10 Dataset for the largest supermarket chains in the UK

In this section, I develop a large scale geospatial dataset for supermarket chains in the UK using multiple data sources. First, filter the properties owned by the leading supermarket chains (Asda, Co-op Food, Iceland, Lidl, Marks

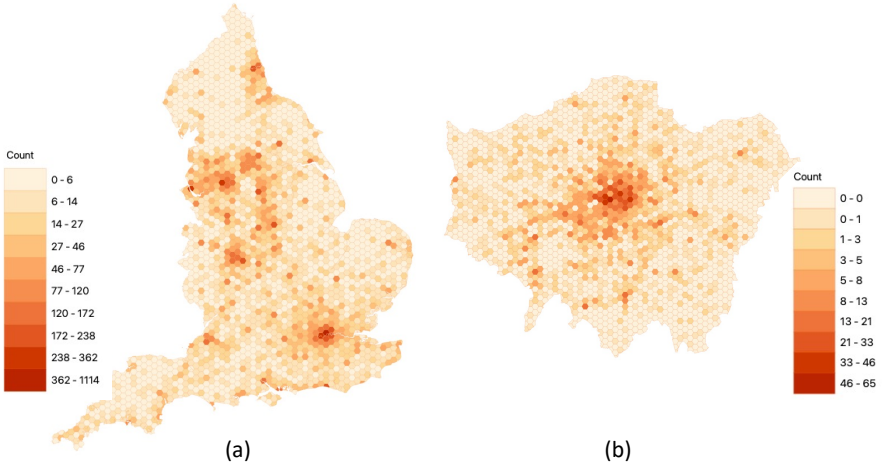


Figure 5.9: Spatial distribution of pubs: (a) across England; (b) zoomed into Greater London. The region is split into equal size grids of hexagons (size of each side : (a) 5km; (b) 0.5km) and number of pubs within each hexagon is displayed with a colour gradient.

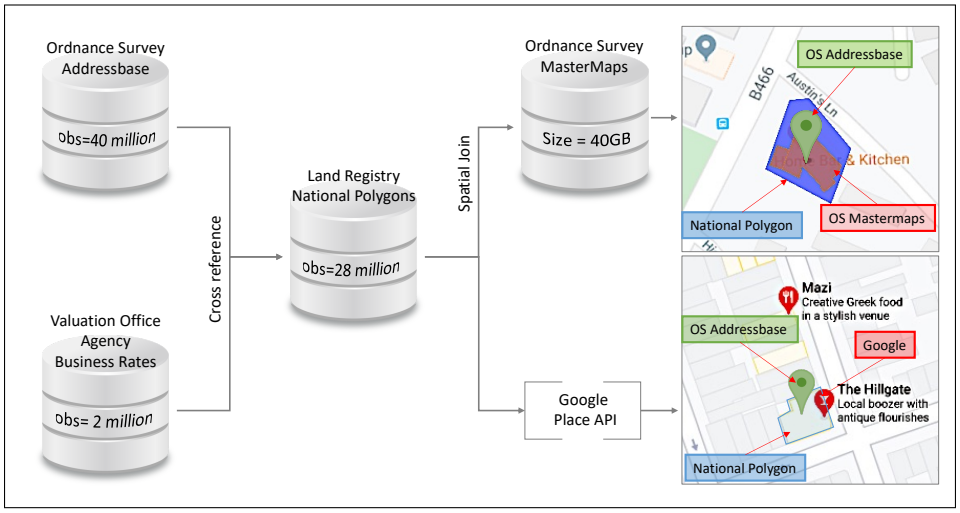


Figure 5.10: Diagram illustrates the steps to extract the store features. Each dataset is named as per the data source along with its number of records (obs) or size. Initially, OS Addressbase 5.2.2 is joined with the non-domestic properties dataset and then spatially joined with National Polygons data to find the Title polygon of each land. This is next joined with Mastermaps and linked with Google data to obtain the store footprints and google customer ratings, respectively.

& Spencer, Morrisons, Sainsbury's, Tesco, Waitrose) from the commercial and corporate ownership data by [73]. However, the filtered data contains other types of businesses owned by the respective supermarket chains, such as their warehouses. [143] data provides the categorisation of the non-domestic properties along with their rateable values and floor sizes. The VOA dataset is filtered to extract the properties representing supermarket or food store

Table 5.4: Summary statistics of the compiled dataset for pubs in London.

	Characteristics	Mean	SD	Min	Median	Max
Pubs internal characteristics	Floorspace (sqm)	287.3	203.7	39.0	238.0	2,499.0
	Height (m)	6.8	3.3	2.5	6.5	20.3
	Number of floors	2.3	1.3	1.0	2.0	8.0
	Total area of land (sqm)	578.2	384.0	44.0	480.0	3,806.0
Distance to the closest (m)	Metro	1,572.9	1,948.7	1.4	730.3	12,481.5
	Train Station	691.3	655.2	7.2	507.4	6,592.7
	Bus Stop	89.0	81.2	3.0	65.6	1,230.2
	Park	1,280.6	1,240.5	2.0	885.3	6,739.9
	Popular Attractions	1,949.5	1,334.5	10.0	1,625.3	6,731.8
	Sports Facility	225.5	176.1	7.2	181.2	1,207.4
Google data	Customer rating	4.2	0.3	1.6	4.2	5.0
	Number of users rated	472.8	473.4	1.0	359.0	5,478.0

categories. Since there is no direct link between the two datasets, VOA data with the OS Addressbase data 5.2.2 are joined with the cross-reference to obtain the geo-coordinates of the properties. Next, the filtered ownership data is joined with the National polygons dataset [72] to identify each properties title polygon. Finally, the two datasets are spatially joined to obtain a dataset of Supermarkets in the UK with their geo-location coordinates. This process is illustrated in a flow diagram such that it is visually easy to read Fig. 5.11. The spatial distribution of supermarkets in Greater London is shown in Fig. 5.12, and the summary statistics of the compiled dataset is presented in Table 5.5.

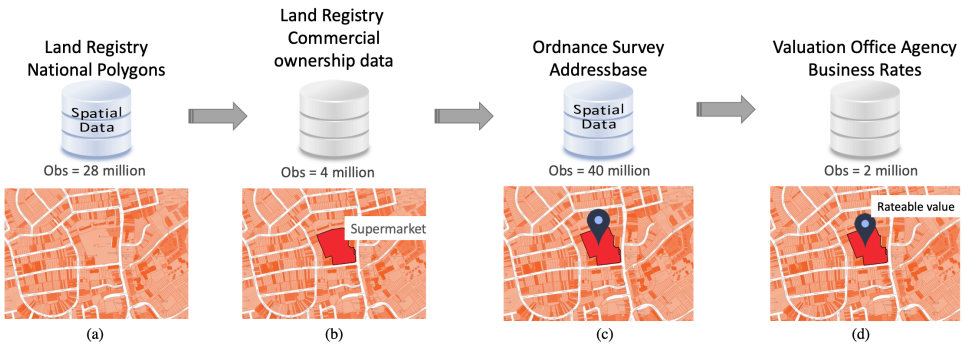
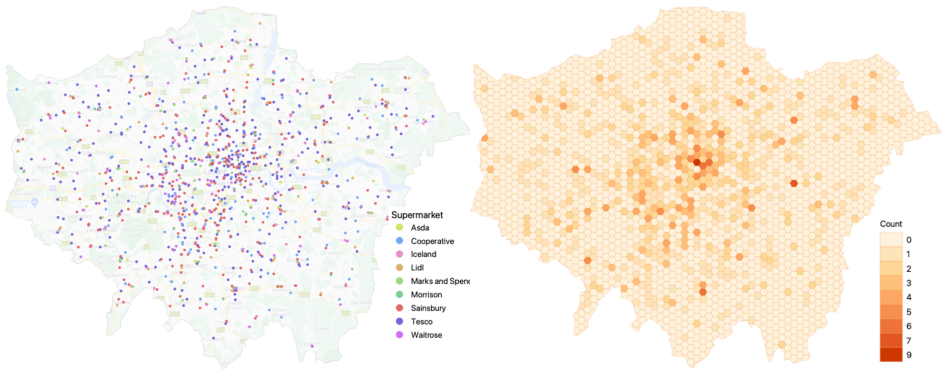
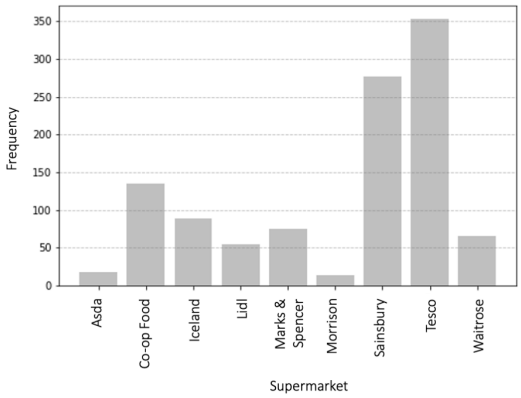


Figure 5.11: The flow diagram presents the steps in developing a dataset of Supermarkets in the UK with their geo-location coordinates. (a) The map shows title polygons. (b) Filter only the titles owned by supermarket chains. (c) Spatially join to identify the data from OS. (b) Finally, join with VOA data to get only the supermarkets and their rateable values.

The annual revenue generated by supermarkets is extracted using the companies' annual statements. Since the individual revenues at each store are not published, I calculate a revenue proxy using the reported annual revenues proportional to their rateable values.



(a) (b)



(c)

Figure 5.12: (a) Visualisation of the supermarket locations with their respective supermarket chains name. (b) Greater London is split into equal size grids of hexagons (size of each side is 0.5km) and number of supermarkets within each hexagon is displayed with a colour gradient. (c) Frequency distribution of the supermarkets

Table 5.5: Summary statistics of the compiled dataset for supermarkets in Greater London.

	Characteristics	**Mean**	**Std**	**Min**	**Median**	**Max**
Store characteristics	Floorspace (sqm)	1,728	2,762	18	518	29,054
Distance to the closest (m)	Metro	1,456	1,793	3	692	9,856
	Train Station	800	730	10	589	6407
	Bus Stop	68	48	2	56	442
	Park	1428	1252	5	1065	6792
	Popular Attractions	2093	1395	7	1794	6665
	Sports Facility	171	172	4	124	1420
Google data	Customer rating	3.7	0.53	1.7	3.9	4.7
	Number of users rated	356	646	2	70	7214

5.3 Customer zone characteristics

In this section, a dataset is formed to use in the application to demonstrate the customer level characteristics. The most granular level of customer data can be identified as the residential locations. OS Addressbase dataset [113] provides both residential and commercial addresses (over 40 million) along with geo-locations. However, since there is no data for customer features at the residential level, in this thesis, I consider postcodes which is the next most granular level. Henceforth, we assume that the customers' behaviour who are residing in the same postcode are homogeneous.

5.3.1 Postcode level data

There are approximately 1.8 million postcodes in the UK, and on average, each postcode has 15 properties. In Greater London, on average, there are 17 households per postcode. The postcode centroids for Greater London are displayed in Fig. 6.3. The dataset can be accessed as point data with the geo-coordinates [107]. Additionally, the postcode population breakdown with gender is available to download as CSV [141]. The population and proportion of gender at the postcode level are extracted to reflect the demographics in the area. This dataset is subject to the Open Government Licence v.3.0.

5.3.2 Indices of Deprivation

In this thesis, the deprivation data are employed to enclose a broad range of an individual's living conditions. For instance, lack of financial resources to support people's needs can be considered living in poverty, but lack of any kind of resources can be considered deprived. Hence there are seven domains of deprivation: (1) Income Deprivation; (2) Employment Deprivation; (3) Education, Skills and Training Deprivation; (4) Health Deprivation and Disability; (5) Crime; (6) Barriers to Housing and Services; (7) Living Environment Deprivation. The Indices of Deprivation [100] provide a set of relative measures of deprivation for small areas (Lower-layer Super Output Areas) across England. This dataset is subject to the open government license. Since the deprivation data is provided at the LSOA level, it is assigned to the postcodes by point to polygon spatial join to match the customer zone data.

5.3.3 Lower-layer Super Output Areas (LSOA)

The Office for National Statistics design LSOA for the purpose of reporting statistics in small areas[106]. Each area consists of similar population size, with an average of approximately 1,500 residents or 650 households. There are 32,844 LSOAs in England. The dataset is available to download in shapefile

format that is consists of polygons. This dataset is subject to the Open Government Licence.

5.4 Summary

In this chapter, the focus was on compiling three main datasets : (1) over 1500 pubs geospatial and related characteristics; (2) supermarket store data for the seven leading chains in the UK; (3) customer zone data at the postcode level. This chapter overcomes one of the major limitations in the literature by presenting real-world data for large-scale experiments. The datasets compiled in this study could be beneficial for future research. All the primary datasets that are available in the property-related industry are discussed in the study. Using the methods introduced opens up many other directions to compile datasets to explore different sectors such as restaurants and cafes.

The next chapter introduces two large scale real-world experiments to explore the methodologies and datasets introduced in this thesis. First explores the non-domestic properties by applying state-of-the-are Fixed ranked kriging. Finally, the BSIM is fitted and subsequently used to identify optimal facility locations for pubs and supermarket industries.

5.4.1 Availability of data

The data that support the findings of this study was generated by combining open source and commercial proprietary data. The commercial proprietary data were obtained from the industrial partner to support research purposes and are subject to strict non-disclosure agreements. However, researchers interested in replicating our results on the commercial problems can directly request data from the relevant organisations using the references made in the paper. The data used for synthetic experiments and data created from the open sources for real-world applications are provided under 'Data' in the GitHub repository available at https://github.com/shanakap/BSIM.

Chapter 6

Real-world Applications

6.1 Introduction

This chapter presents multiple real-world applications to demonstrate the methodologies introduced by employing the large-scale datasets introduced in the previous chapter. I initially explore the spatial variations in non-domestic properties rateable values by applying the Kriging method. The following sections present two applications for a subset of non-domestic properties: pubs and supermarkets. These two applications demonstrate the proposed BSIM method and apply it to make location decisions to enter a new market or expand in an existing competitive market.

My main contributions are: (a) application of Kriging to explore spatial variation in rateable values across different categories of non-domestic properties; (b) BSIM method is applied to pubs and supermarket sector that proved to provide the best predictive performance compared to competing approaches while providing inference at the level of customers and business facilities, delivering invaluable insights for planning and decision making; (c) demonstrated the optimal facility locations and their designs for a new company to enter the pubs' industry and expand the existence of a supermarket chain in Greater London.

To the best of my knowledge, I am the first to present a fully integrated competitive facility problem that includes both the spatial interaction modelling component and the store location optimisation framework demonstrated in one of the major cities in the world using a large-scale dataset with over 1000 supermarkets, 1500 pubs, and 150000 customer regions.

6.2 Modelling Business Rates with FRK

The first application is primarily interested in exploring the non-domestic dataset presented in Chapter 5. In this application, I propose a state-of-the-art

Fixed Rank Kriging model to cope with high-dimensionality and learn rateable values from the spatial context and category of the non-domestic properties. By accounting for spatial effects, the model improves current business rates valuation practice and could help in making the process more fair and transparent.

Three different formulations are evaluated in modelling the logarithm of rateable values:

1. Model with no covariates.
   This model only uses the spatial coordinates to fit the model and will not use the information about the category of the properties.

2. Model for each category with no covariates.
   An individual model fits for each category using spatial coordinates.

3. Model with category as the covariate.
   The category of the non-domestic property and spatial coordinates are used in the model.

6.2.1 Cross validation

The standard data sampling methods used for cross-validation (CV) to evaluate prediction performance assumes the training and testing data are independent of each other. According to the first law of geography, "Everything is related to everything else, but near things are more related than distant things" [139]. This causes the standard sampling methods to produce optimistic performance measures for spatial models. Spatial k-fold cross-validation (SKCV) is a modification method of the standard CV to remove the spatial autocorrelation (SAC) between the training and testing data [119]. This is achieved by removing training data within a pre-determined radius, known as the deadzone, around the test data. There is a trade-off between the radius of deadzone and the loss of data in the training sample. Three data sampling methods used for CV in this study:

1. Standard k-fold CV

2. SKCV with 20m deadzone

3. SKCV with 50m deadzone

6.2.2 Results

The three model variations discussed are fitted for each of the cross-validation methods. Under the k-fold sampling 90% of the data were used in training at each fold, but for SKCV on average, only 68% and 38% of the data were used

for 20m and 50m deadzones respectively. Further increase in deadzone radius would result in less data for training; hence 20m and 50m radius were used for cross-validation.

FRK predictions are obtained for point locations on the test data set for each fold. Fig. 6.1 shows the outcome of the three models for predictions at all points using SKCV with 20m deadzone. Similar patterns were observed for the other sampling methods. Hot spots of high rateable values are observed in the centre of each map, which represents Central London. 6.1(a) is more smoother compared to 6.1(b) and (c). This is likely due to the fact that Model 1 is not using the category of the property in modelling.

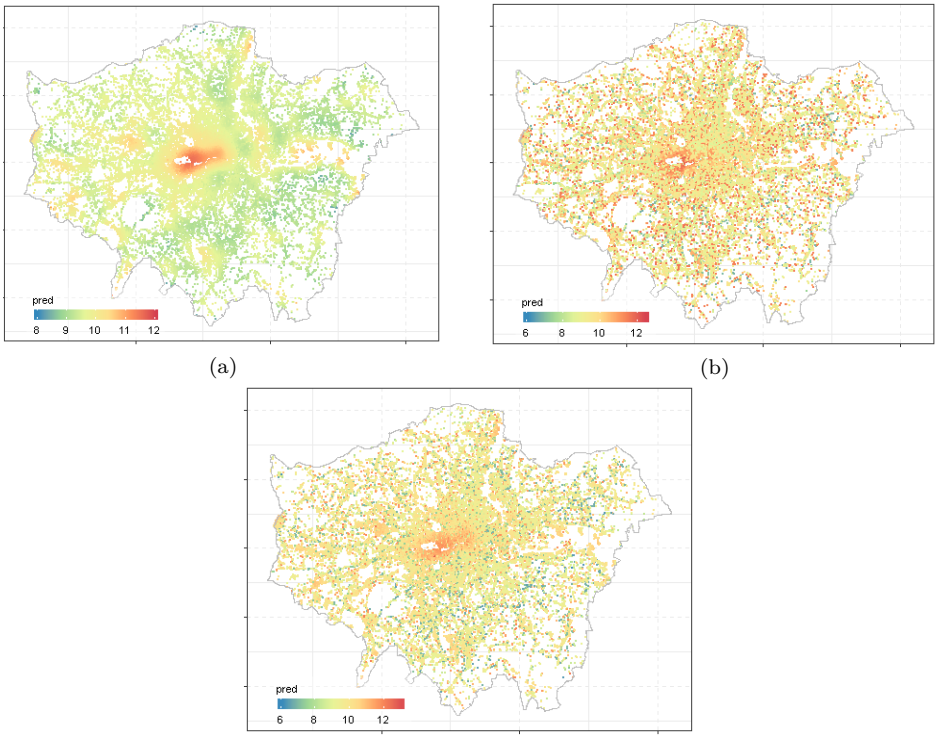


Figure 6.1: Prediction of log Rateable value obtained from FRK: (a)Model 1; (b)Model 2; (c)Model 3.

A summary of all the performance for validation data are recorded in table 6.1. The bold font represents the best model for each sampling technique. R^2 double when the model uses information about the category of the property. This emphasises that business rates are influenced significantly by the category of the property in addition to the location. SKCV with a deadzone is utilised to penalise the over bias caused by spatial autocorrelation. The k-fold cross-validation is providing optimism due to the overestimation of statistical effects, but 50m deadzone removes 60% of the training set, so although it removes the SAC between the training and test set, it also provides pessimism in the fact that it has a smaller training set. However, notably, there is no significant

difference in the performance across the sampling techniques.

Table 6.1: Results table for three models with the three validation techniques.

	k-fold			Dead zone - 20M			Dead zone - 50M		
Model	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
R^2	0.17	**0.54**	0.49	0.16	**0.53**	0.49	0.14	**0.52**	0.47
RMSE	0.096	**0.072**	0.075	0.097	**0.073**	0.078	0.098	**0.074**	0.079
MAPE	9.48	**6.96**	7.27	9.56	**7.04**	7.37	9.7	**7.13**	7.49

In order to understand the performance of the model, for each category, I have calculated the R^2 of the validation data for the three models and shown in Fig. 6.2. Model 2 performs better for each category compared to models 1 and 3. There is a notable difference in R^2 for three models in the *sales kiosks* category, which has the least number of properties in the subset used for London (Fig. 5.3). Furthermore, *restaurants* shows the highest R^2 under all the three models despite representing only 2.5% of the data. Overall Fig. 6.2 shows that the predictability varies greatly (R^2 between 0.01 to 0.48) with the category of the property and in combination with Fig. 5.3 provides no evidence this is driven by the number of observations in each category. Furthermore, *restaurants*, *cafes* and *offices* show similar R^2 for all three models which indicates that these categories are representative of the overall system compared to *Sales Kiosks*, *Pubs*, *Business units* and *Factories* tend to have their own subsystems.

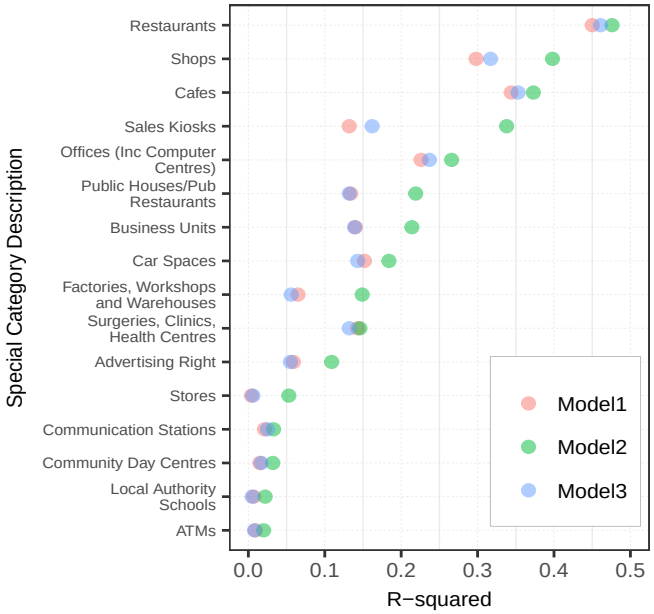


Figure 6.2: R^2 for each property category in SKCV with 50m deadzone.

6.2.3 Limitations

There are various limitations in applying kriging for interpolation. The accuracy is limited if the number of observations is small, data has restricted spatial coverage, or the data is not sufficiently correlated. The interpolated points have to lie within the range of the observations. Hence if the peaks are not sampled, then they cannot be inferred accurately. In the study, this is made more concerning when using the spatial k-fold cross-validation method as it can ignore the peaks when fitting the model, leading to poor results. The interpolated points have to lie within the range of the observations. Hence if the peaks are not sampled, then they cannot be inferred accurately. In the study, this is made more concerning when using the spatial k-fold cross-validation method as it can ignore the peaks when fitting the model, leading to poor results.

Furthermore, kriging cannot account for additional dimensions such as the temporal variation in the data. Hence the method is limited to only predicting the rateable values in a certain time horizon. However, the predictive rateable value model needs to have the possibility to forecast as the values change over time. Additionally, the study is limited to Greater London, drawing a hard boundary and ignoring the observations outside the area. Hence the data beyond the boundary are not reflected in the model. Therefore the estimation of the values at the edge could perform less accurately.

The correlation between the different sectors is not accounted for in the kriging model. However, the Co-Kriging method can be adapted to predict the unobserved locations by combining known spatial attributes and correlated variables or simultaneously predicting values of two or more sectors [104]. The limitation with such a method is that it doesn't scale up well, such as the method used in this study, FRK.

This study models the total rateable value with the FRK method. One of the key limitations is that the size of the non-domestic property is not accounted for in model fitting. The VOA data does not provide all industries' total building sizes, and it is a tedious process to acquire that data. Additionally, there are sectors such as public houses where the rateable value is not decided on the internal area but is calculated based on their revenue. Henceforth the results may be improved by considering the various valuation schemes used to calculate the rateable value by the Land Registry.

6.3 Case study 1: Optimal locations for entering into a market concerning the pub industry

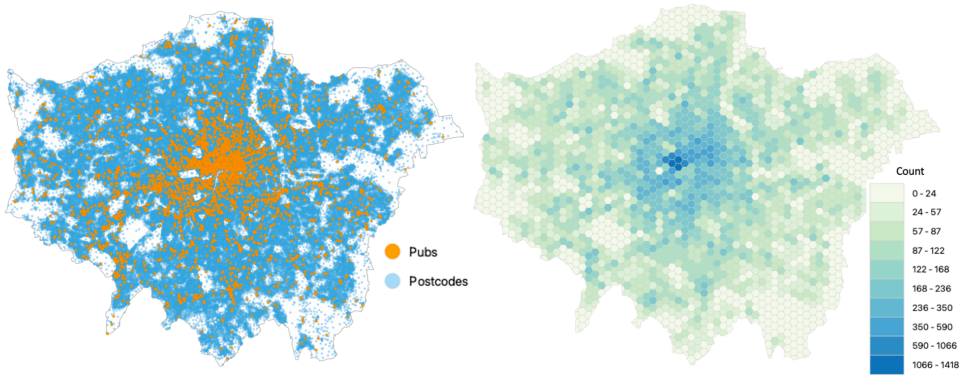
In this section, I illustrate our proposed methodology using the pubs' dataset developed for Greater London in Chapter 5. After compiling data from different sources, the final complete dataset consists of $S = 1804$ pubs. Therefore some pubs are not considered in the study, hence unable to capture the competition accurately. The derived approximated revenue after adjusting for edge correction (Eq. (3.28)) is used as the response variable y_s in the model with natural log transformation. The approximated revenue may not provide accurate predictions. For each pub, specific features are derived: floorspace, height, number of floors, the total area of land; distance to the closest metro, train station, bus stop, park, popular attractions, sports facility; customer rating on Google, number of users rated and an indicator to show if the pub is in a major town. The Euclidean distance is considered when creating spatial features such as the distance to public transport access points. However, the use of a road network would provide more realistic results. Additionally, the number of users rating the pubs would depend on the period of its operations. Hence a more realistic approach is to normalise the value based on the period that each facility was in operation.

The customer locations are considered to be at the postcode level, which is the most granular level of census estimates are released. There are $N = 174360$ postcodes for Greater London. The characteristics of the postcodes are represented by the population at each postcode and its proportion of male, and deprivation scores. All features have been normalised before training the model. People living in the postcodes are assumed to have similar behaviour, but this may not hold in reality. The model may be improved with more granular customer-specific characteristics; underlying arguments would remain the same. Centroids of the postcodes and retail locations of the pubs are presented in Fig. 6.3(a), on a map of London.

6.3.1 Estimating revenues using the BSIM

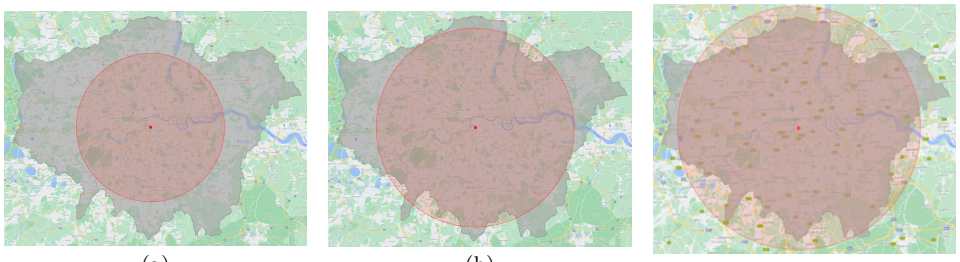
Customer behaviour is not affected after a certain distance from the business facility, despite the pubs' attractiveness. The model is explored under three different radius, $d_T = 15$km, 20km and 25km as presented in Fig. 6.4. The distance between origin and pub is calculated using Euclidean distance, although a better representation would use a transport network. Hence the maximum distance a customer travelling cannot be accurately determined by the radii selected to truncate the Gaussian.

I first perform a preliminary study of our model with a store-specific coeffi-



(a) (b)

Figure 6.3: (a) Visualization of the locations of pubs in orange markers ($S = 1804$) and postcode centroids in blue markers ($N = 174360$) over the map of London; (b) Greater London is split into equal size grids of hexagons (size of each side is 0.5km) and number of postcodes within each hexagon is displayed with a colour gradient.



(a) (b) (c)

Figure 6.4: Demonstration of different radius used for truncated Gaussian with an example concerning a pub located in the center of London. Three radii were used in the study: (a) 15km; (b) 20km; (c) 25km.

cient which denotes the store-specific variance $\sigma_s^2 = \exp(\upsilon_s)$, representing the attractiveness of the store as given by Eq. (3.1). The model is experimented with three different radii of the truncated Gaussian and model performance summarised in Table 6.2. Results indicate that R^2 increased to 0.72 as the radius increased from 15km to 20km but reduced to 0.57 as the radius increased to 25km. Hence the best experimental results yielded for truncated Gaussian with a radius of 20km.

Table 6.2: R^2, γ^{-1}, NRMSE and coverage for the fitted BSIM with revenues of pubs in Greater London under three different radii of the truncated Gaussian.

Truncated radius (km)	R^2	γ^{-1}	NRMSE	Coverage
15	0.19	0.67	0.08	94%
20	0.72	0.45	0.05	96%
25	0.57	0.52	0.06	95%

Next, a detailed study is performed on the model with improved specifica-

Table 6.3: Coefficient (λ) of the store features† for the best fitted model.

Variable	**coefficient**	**CI (95%)**
Floor space	-0.14	[-1.21, 0.92]
Height	0.22	[-0.76, 1.19]
Total area	0.13	[-1.08, 1.34]
Distance to metro	-0.89	[-1.75, -0.03]
Distance to rail	-0.04	[-1.01, 0.94]
Distance to bus stop	-0.10	[-1.3, 1.09]
Distance to closest park	0.31	[-0.46, 1.08]
Distance to closest monument	0.47	[-0.39, 1.33]
Distance to closest sports	0.37	[-0.47, 1.21]
Google customer rating	0.27	[-0.42, 0.95]
Number of users rated	0.70	[-0.46, 1.86]
Store in Town	0.44	[-0.03, 0.91]

† All the variables are normalised before using in the model.

tions where store features represent the attractiveness of the store (Eq. (3.2)). In this study, the radius of the truncated Gaussian is set to 20km, as it demonstrated the best results for the previous experiment. The model with these settings resulted in a high R^2 of 0.88, coverage over 96% for the 95% CI and a low NRMSE of 0.03. The plots (Fig. 6.5) of the observed revenue and predicted revenues suggest that the model provides a good fit for the data. It is important to consider that the evaluation is for in-sample data, but a better approach would be to evaluate against out-of-sample data for testing. The cross-validation techniques cannot be applied because all the store data are required to model competition at a certain time horizon. Hence, this study is restricted to in-sample because of the limited access to temporal data. The residuals are relatively high out of central London, closer to a major ringway as shown in Fig. 6.5 c. Additionally, the predicted revenue of a few pubs at the edge of the study area is underestimated, reflecting the edge effects. This could be because in the edge correction the competition of the pubs outside the study area is not adjusted.

Using the parameter estimates (λ, ε) from the best-fitted model, the attractiveness (σ_s^2) of pubs are demonstrated around London in Fig. 6.6(a). It can be observed that the most attractive pubs are within or around the major towns. Further exploring the coefficients (λ) for the best-fitted model presented in Table 6.3, it is observed that number of people rated on Google had the highest positive contribution towards the attraction term. This implies that customer rating is a critical indicator in describing the customer attractiveness to the pubs. One could argue that a higher number of customers rated the pub as it was operating for a longer period. Hence there have to be more variables, such as the start dates of the pubs, to conclude this remark accurately.

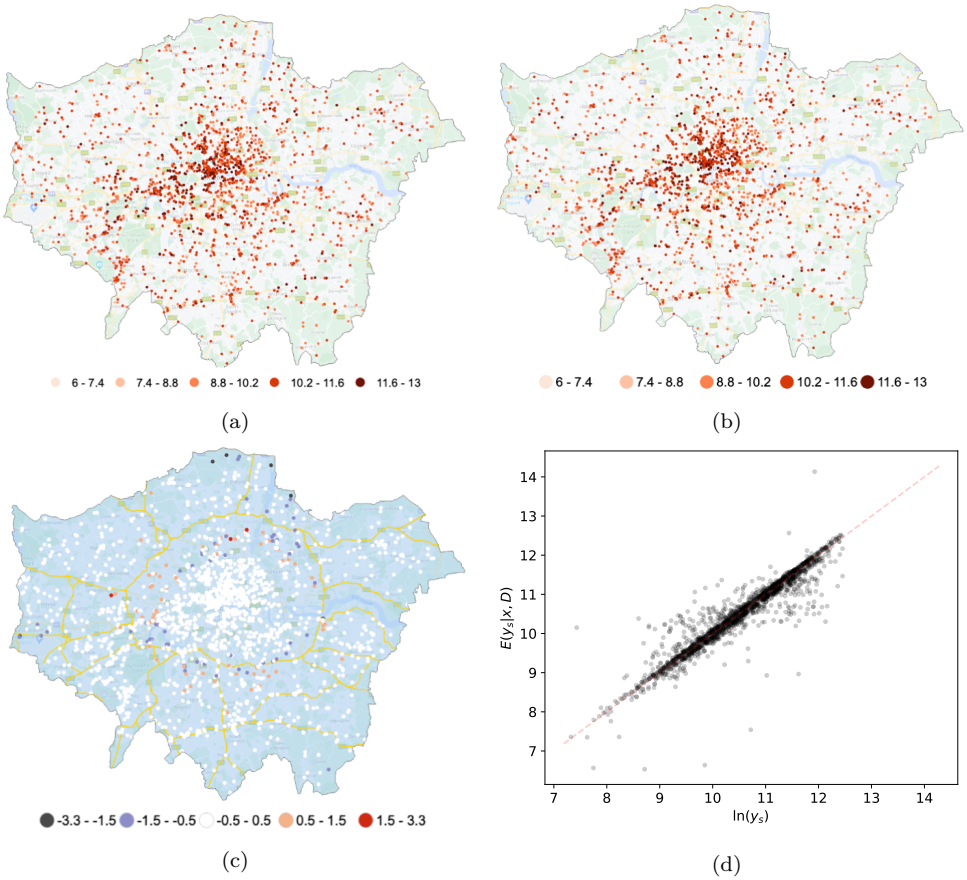
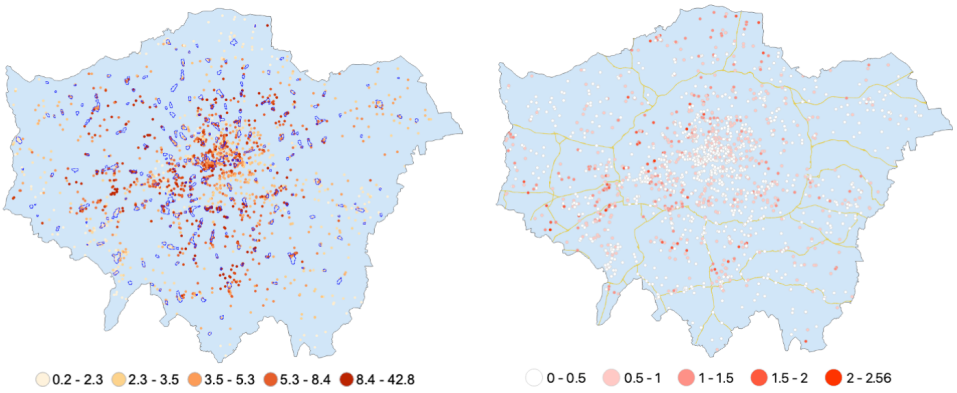


Figure 6.5: Visualisation of the Pub's revenue and predictions over greater London map with truncated Gaussian radius of 20 km: (a) Revenue at each pub; (b) Predicted revenue at each pub; (c) Residuals marked in points and lines are the major roads; (d) Actual against predicted revenue. The experiment resulted in $R^2 = 0.88$ and NRMSE = 0.03.

The remaining term that expresses the attractiveness, unobserved pub features (ε_s), where the absolute coefficient is mapped in Fig. 6.6(b). There is a similar pattern to the residual plot, but the overall spatial distribution appears to be random. A deep investigation is required to understand what could explain the unobserved pub features.

For demonstration purposes, a pub in central London is randomly selected to explore the insights from the fitted model. The probability of people within the postcode selecting the particular pub (p_{ns}) is calculated using the model parameter estimates with Eq. (3.4). These probabilities are mapped into a heatmap as shown in Fig. 6.7. There appear to be two hotspots on the map, one closer to the pub, and another one towards North-West London. It is natural to see higher probabilities closer to the pub, but the other hotspot is possible because the pubs' density in the area is comparatively low, as shown in Fig. 6.3(a). Hence people in the area also prefer travelling to pubs in central London. The distribution of probabilities tends to be having an



(a) (b)

Figure 6.6: Exploring the pubs attractiveness for the fitted model: (a) Variance (σ_s^2) of the Gaussian placed on each pub. Blue colour polygons denote the major towns; (b) Absolute coefficients of the unobserved pub characteristics (ε_s).

oval shape, possibly because the distance between the customers and pubs is calculated as Euclidean distance. Furthermore, the pub with the highest p_{ns} can be regarded as the most likely pub to visit by the people living in a given postcode. This interpretation is extended to infer the cluster of postcodes with similar preferences and presented as a motivational example in Chapter 1.2.

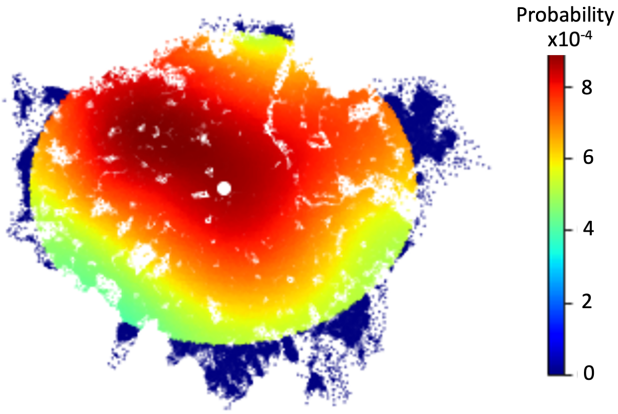


Figure 6.7: Visualisation of the probability (p_{ns}) of people in each postcode selecting the particular pub shown in a white dot in the centre of London.

Coefficients of the deprivation features are presented in Table 6.4. The results indicate that areas with higher income, high employment, less risk of crimes, a better quality of life, and the environment positively influence the customers' spending at pubs.

The amount spent by customers living in each postcode (b_n) can be estimated using the best-fitted model's parameter estimates ($\boldsymbol{\beta}$). The amount spent at each Borough can be derived by calculating the total of the estimated spend-

Table 6.4: Coefficient (β) of the customer features † for the best fitted model.

var	coefficient	CI (95%)
Postcode Population	-0.19	[-1.26, 0.87]
Male Proportion	-0.18	[-1.15, 0.79]
Income Deprivation	0.19	[-1.02, 1.4]
Employment Deprivation	0.22	[-0.64, 1.08]
Education, Skills and Training Deprivation	-0.51	[-1.49, 0.46]
Health Deprivation and Disability	0.46	[-0.74, 1.65]
Crime Deprivation	0.32	[-0.45, 1.09]
Barriers to Housing and Services	-0.28	[-1.14, 0.58]
Living Environment Deprivation	0.26	[-0.59, 1.1]

† All the variables are normalised and one minus the normalised deprivation value is obtained to reflect higher values as better areas. For example higher value for transformed income deprivation would mean a wealthy area.

ing amount at each postcode within the Borough. This is compared against the alcohol-related mortality in the London Boroughs published by [120]. The rank of Boroughs respective to the spending and mortality levels published for 2017 is mapped in Fig. 6.8. The rank correlation between mortality count and estimated spending shows a moderate positive relationship of 0.4. Our intuition is that higher alcohol-related mortalities are to be expected in the areas of high alcohol consumption. However, it is important to consider that the formulation of budgeted spending is based on the assumption of a linear relationship with customer characteristics. Hence, this assumption should be verified against actual data before making final conclusions. Furthermore, the Fig. 6.8 (a) shows that central London is top of the ranking for spending, but the mortality ranking is lower. This could be an indication that the model is unable to capture the high demand generated in central London from the customers living outside London.

Finally, a comparison is performed with a spatial interaction model from the literature for completeness of the study. The Modified Huff model [95] is fitted for the same dataset, which displayed very low performance with R^2 of only 0.03 and NRMSE of 0.84. Our model outperforms the benchmark model with a notable improvement and provides valuable inferences for decision-makers.

6.3.2 Optimal facility locations

In this application, the data and parameter estimates evaluated in the previous spatial interactions study is applied. Optimisation problem considers three sizes of pub designs with total floor size: 175 sqm, 500 sqm and 1275 sqm. The cost of each design is calculated based on the ratios between the sizes: 1, 3 and 7 are the costs of constructing small, medium and large size pubs, respectively. For simplicity the cost of constructing a pub anywhere in London is considered to be eqaul but in practice the cost of construction in

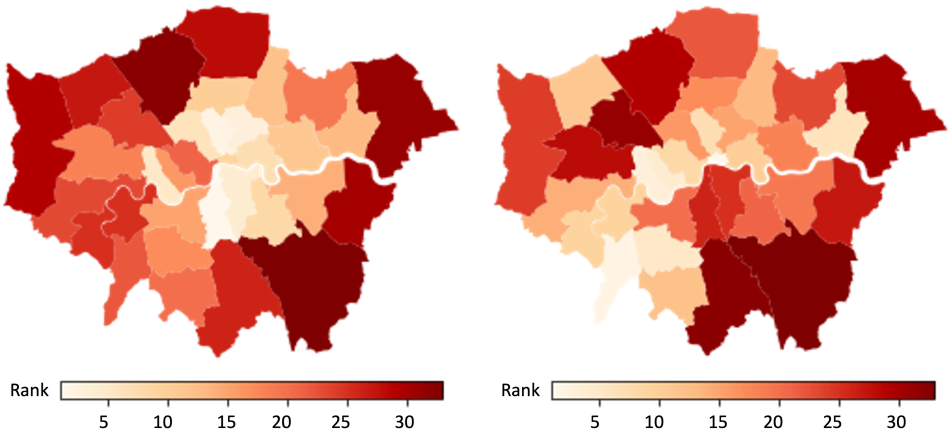


Figure 6.8: Visualisation of ranking on estimated revenue and mortality in the London Boroughs : (a) Rank of estimated amount spent at the pubs by people living in each Borough; (b) Rank of Mortality count.

central London is significanly higher compared to rest of the area. Additionally, the other characteristics for each potential facility location are calculated as discussed in Chapter 5.2. Google ratings at the new sites are assumed to achieve the average ratings of the existing pubs. There could be more insights in the Google ratings that could be relate to the new facilities such as the pub size and region. This could improve the estimation of Google rating rather than using the average across all new pubs. The unobserved pub features (ε_s) are assumed to be similar to the average of existing pubs within the grids used for sampling new potential locations.

I search for optimal locations to build at most two pubs using a budget of nine. The multiresolution sampling method is used with 5×5 grids with a depth of three to generate the initial set of candidate locations. The generated potential locations are split into four random samples and executed the optimisation algorithm parallelly. Two optimal locations for a new company entering the pubs market is detected to be in Redbridge and Bromley with small and large structures, respectively, as demonstrated in Fig. 6.9(b). The optimal pub locations are tend to be in the edge of the study area, this could be mainly because the model is unable to capture the competition created by the pubs outside the study area. Henceforth there has to be a method to capture the competition on edge before making the final decision on the optimal sites.

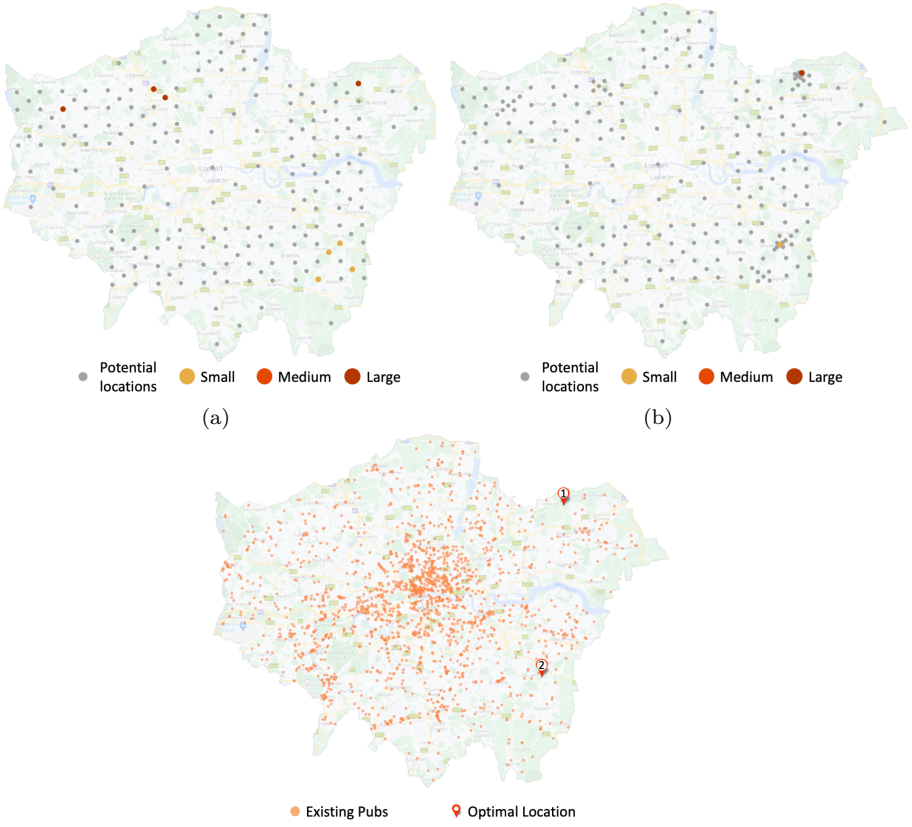


Figure 6.9: Optimal sites to establish two pubs in London. (a) Optimal locations from four independent samples. (b) All the potential locations that were eventuated at different stages and the final optimal locations. (c) Existing pubs and new optimal facility location.

The median and 50% credible interval(CI) of the estimated revenue for the two locations are displayed in the Table. 6.5. The monthly estimated sales of both the pubs are higher than the average revenue generated by the existing pubs within their respective boroughs. It is important to consider that the revenue in the study is approximated using the rateable values and not the actual turnover of the businesses. Hence the experiments need to be conducted with the actual revenue to make final conclusions.

Distance between the optimal locations and the public transport access points and key venues are presented in Table. 6.6. New sites are located near sports facilities and closer to bus stops. The distance is calculated as the Euclidean distance and not the actual road distance. Hence the final results may change when adopting the road network. Revenue at the small pub is expected to be driven by the customers attracted to the area with key venues.

The area of the small pub is explored with an eagle view in Fig. 6.10. A park, monument, and gymnastics centre are located near the optimal location, meaning a busy area with more people interactions. There is only one pub

Table 6.5: Monthly revenue† estimations of the two optimal pubs reported in millions

Pub	Borough	Average Revenue in the Borough	Estimated Revenue	
			Median	50% CI
1	Redbridge	0.62	1.25	(0.75, 1.88)
2	Bromley	0.54	4.26	(2.52,10.67)

†Revenue at the existing pubs are derived using the business rateable values.

Table 6.6: Characteristics of the two optimal pubs

Pub	Design	Floor size (sqm)	Distance to the nearest (m)					
			Metro	Rail	Bus	Parks	Attractions	Sports Facility
1	Small	175	8,36	2,136	533	312	224	678
2	Large	1,275	3,521	4,564	733	6,206	4,285	718

within the 1 km radius, indicating less competition for the new pub. A bus stop is located within walking distance, offering people easy accessibility to the location. However, there is no main road access to the site, thus including distance to the main road as a store feature could provide more realistic results.

Figure 6.10: Eagle view of the optimal pub location with the small design. The dashed squares indicate some of the key venues in the surrounding of 1 km radius.

6.4 Case study 2: Best sites to expand a chain related to Supermarkets

In this section, first, I demonstrate the extended BSIM by modelling the revenue of supermarkets and subsequently find the optimal locations for a supermarket chain to expand in Greater London.

This case study is centred on the seven largest supermarket chains in the UK. The complete dataset consists of $S = 1079$ supermarkets located within Greater London. The derived store features are used for each supermarket store: floorspace, customer rating on Google, number of users rated, an indicator to show if the supermarket is in a major town and distance to the nearest metro, train station, bus stop, park, popular attractions, sports facility. The postcode level data represents customer locations and their characteristics: population, the proportion of males, and deprivation scores. As discussed in the previous case study, this real-world experiment also inherits the limitations in the features used.

6.4.1 Estimating revenues using the extended BSIM

The BSIM parameters are estimated under four truncated radii for the Gaussian distribution and summarise the performance in Table. 6.7. The reasonable distance a customer is willing to travel is assumed to be half of the maximum extent a customer would travel ($d_D = d_T/2$). The results demonstrate that R^2 increased to 0.89 while increasing the truncated radius to 20 km. However, R^2 decreased significantly when it reached a 25 km radius that covers the whole of London. Similar results were obtained in the BSIM study for pubs in the previous section.

Table 6.7: R^2, γ^{-1} and NRMSE for the fitted extended BSIM for revenues of supermarkets in London under four different radii of the truncated Gaussian distribution

	Truncated radius (km)			
	10	15	20	25
R^2	0.38	0.64	**0.89**	0.1
γ^{-1}	0.64	0.5	0.31	0.72
NRMSE	0.07	0.05	0.03	0.09

The results from the best-fitted model are demonstrated in Fig. 6.11. The scatter plot with the actual log revenue against the predicted log revenue in Fig. 6.11(a) shows that there are predicted values with large deviance from the actual in tails of the distribution. The spatial distribution of the predicted values are shown in Fig. 6.11(b). The residual values are explored for each

supermarket chain in Fig. 6.11(c). Tesco and Sainsbury's, the two chains with the highest number of stores, 353 and 277 respectively, show larger variance for residual values. The spatial distribution of the residuals exhibits to be randomly distributed, as shown in Fig. 6.11(d).

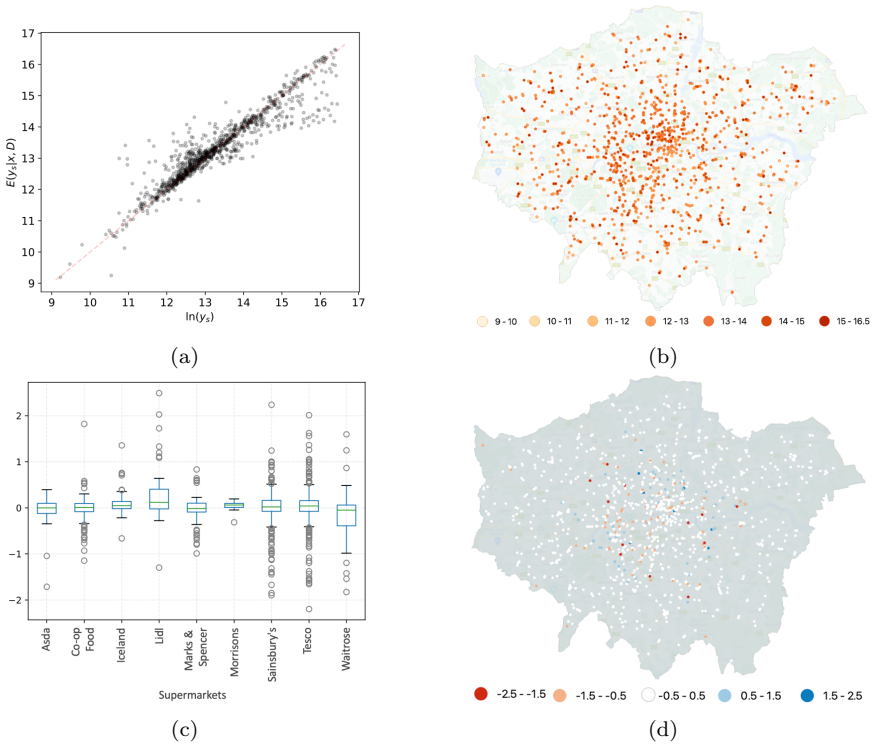


Figure 6.11: The results of the best-performed experiment for the extended BSIM with the supermarkets' revenue. (a) Actual against predicted revenue at each supermarket. (b) Predicted revenue at each store. (c) Residuals against the supermarket chain. (d) Spatial distribution of the residuals.

6.4.2 Optimal facility locations

The parameter estimates from the best-fitted extended BSIM are used to calculate the objective function of the optimisation problem. There are four types of supermarket stores with varying floorspace: Express (278 sqm), Metro (1021 sqm), Superstores (3251 sqm), Extra (5574 sqm). These sizes are used as the possible designs to structure the new facilities. The cost of each design is calculated based on the ratios between the sizes: 1, 4, 12 and 20 for constructing Express, Metro, Superstores and Extra stores, respectively. The cost of building a supermarket across Greater London is assumed to be constant, but this will not hold in practice. Additionally, parking for supermarkets is an essential factor that is very challenging to find in central London. Hence these aspects should be considered in making the final decisions.

Additionally, the other characteristics at each potential facility are calcu-

lated, and for Google ratings, it is assumed to have the average ratings of the existing stores for each chain. Optimal locations to build at most two supermarkets are evaluated within a budget of 35. The optimal locations are demonstrated for the largest supermarket chain in the UK, Tesco.

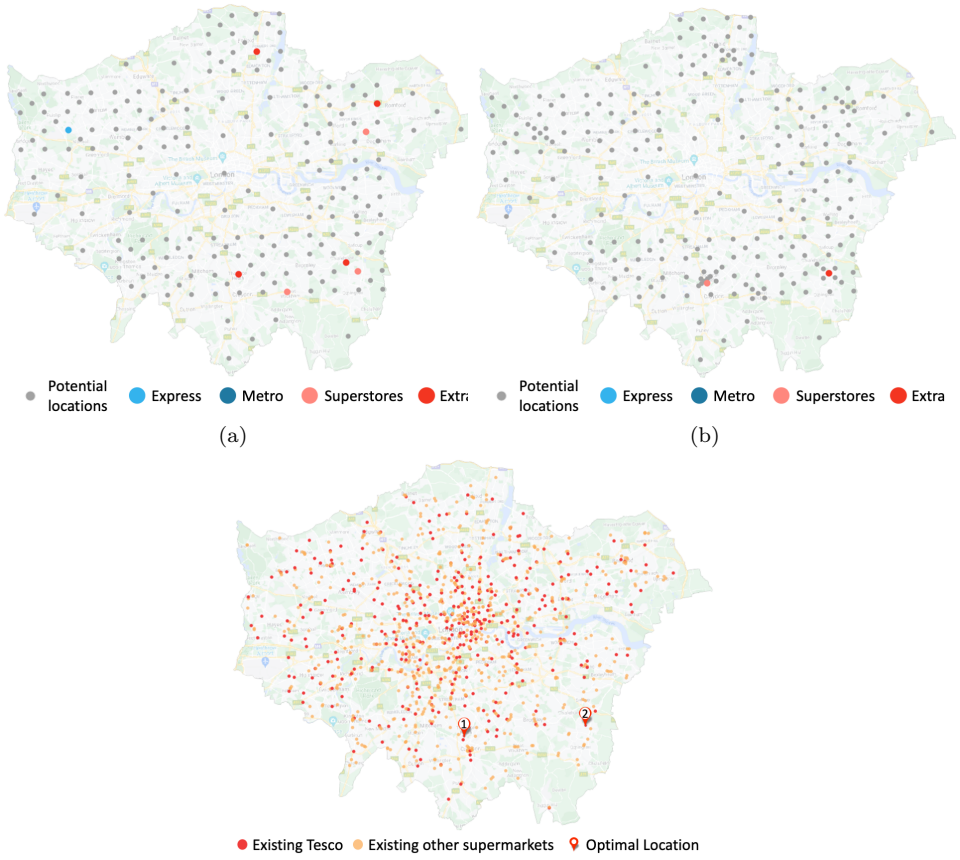


Figure 6.12: Optimal locations to establish two Tesco supermarkets. (a) The initial set of candidate locations is generated from multiresolution sampling with 5×5 grids with a depth of three and optimal locations from four independent samples. (b) All the potential locations that were evaluated at different stages and the final optimal locations. (c) Existing Tesco and other supermarkets and new optimal stores.

The supermarket chains search for optimal locations not just to optimise the revenue at the new facility but to have less impact on the revenues generated at their existing facilities. Hence Eq. 4.16 is used as the objective function. The multiresolution sampling method is used with 5×5 grids with a depth of three to generate the initial set of candidate locations. The generated potential locations are split into four random samples and executed the optimisation algorithm parallelly. Four different optimal sets of locations are detected with varying facility designs as displayed in Fig. 6.12(a). The search algorithm continued for two iterations evaluating the neighbourhood locations produced from quadtree. The optimal locations for Tesco supermar-

ket are detected to be in Croydon and Bromley with designs of a Superstore and Extra respectively as shown in Fig. 6.12(b). No new facility is to be located in the same area as one of the competitors. Similar to the previous case study, the optimal locations for the supermarkets are away from the centre. This is mainly because the model is not accounted for the competition created by the supermarkets outside greater London.

The median and 50% credible interval (CI) of the estimated revenue for the two optimal supermarkets are reported in the Table. 6.8. Both the facilities are predicted to generate more revenue than the average revenue produced by the existing supermarkets in their respective Boroughs. It is important to regard that the revenue is an approximated value and may not provide accurate estimations. The recommended new supermarket in Bromley is located in a less dense area as shown in Fig. 6.12(c). There is only one Tesco and 16 other supermarkets in a 5 km radius (Table. 6.9), compared to the average of 40 Tesco and 59 other supermarket chains found around the existing Tesco supermarkets in London. Significantly high predicted revenue and less competitive location demonstrate an ideal site for a new Tesco store. There has to be more investigation, such as identifying the availability of commercial properties for development and access to the site before making the final decisions.

Table 6.8: Monthly revenue estimations of the optimal stores reported in millions

Supermarket	Borough	Average Revenue in the Borough	Estimated Revenue	
			Median	50% CI
1	Croydon	1.3	2.7	(2.2, 3.5)
2	Bromley	1.6	6.7	(5.3, 8.4)

Table 6.9: Characteristics of the two optimal supermarkets

Supermarket	Design	Floor size (sqm)	Stores in 5km radius †		Distance to the nearest (m)		
			Tesco	Others	Rail	Bus	Sports Facility
1	Superstore	3,251	18	43	1,144	200	791
2	Extra	5,574	1	16	309	57	631

† On average there are 40 Tesco and 59 other supermarket chains around 5 km radius of the existing Tesco supermarkets in Greater London.

6.5 Summary

In this chapter, I have presented three real-world applications: (1) fixed ranked Kriging to model the business rateable values; (2) applied BSIM to model revenue of the pubs in Greater London and thereby search for the optimal locations and corresponding designs for a newcomer to enter the market; (3) seven leading supermarkets revenues are modelled using the extended BSIM and thereby search for best sites for a chain to expand their existence. I have demonstrated how BSIM outperforms competing approaches by evaluating the case studies in terms of prediction performances while providing results that are both interpretable and consistent with related indicators observed for the London region. The introduced modelling frameworks in Chapter 3 and 4 are proven to provide valuable insights for planning and decision-making in the real-world context under uncertainty.

However, it is essential to understand the limitations in interpreting the results of the real-world case studies. As in many other studies, all the factors are not accounted for in the models or experiments. Various attributes such as the social, political and economic climate are not precisely accounted for when modelling revenue or making the optimal location. Hence before making the final decisions, the experts in the respective sectors should be consulted.

The edge correction technique is only applied to address the customers outside the study area but does not account for the competition created by the stores outside the area. This can be observed in the optimal location study where the sites are identified away from the centre. Additionally, road access to the potential sites is not considered in exploring the optimal sites. Hence the sites identified in the study may not be suitable for practice before making the necessary adjustments to the data.

The BSIM is only evaluated for cross-sectional data at a certain time with the real-world data. Hence due to the limited access to data, the forecasting capacity of the model is yet to be validated. Therefore the forecasted revenue for the optimal locations may not provide accurate predictions. The modelling and case studies have presented an advanced approach to spatial interaction modelling in practice, but there are various elements to make it better.

The next chapter makes concluding remarks while discussing the limitations and future extensions to improve the modelling frameworks and experiments presented in the thesis.

Chapter 7

Conclusions and Future Work

The work in this thesis is motivated by the requirement to quantify uncertainty in formulating mathematical models for making location decisions in urban environments. More specifically, the thesis focuses on developing a state-of-the-art mathematical framework to identify the optimal location for businesses under uncertainty by accounting for the underlying spatial interactions in urban systems. Henceforth, build upon the Bayesian framework for uncertainty quantification due to the practical and philosophical reasons discussed in Section 2.4. The thesis provides several new methods that improve the existing modelling capabilities on socio-economic systems and demonstrates their applications in real-world problems. A discussion and conclusions of the work in this thesis are presented in Section 7.1 and further extensions are identified in Section 7.3.

7.1 Discussion and conclusions

Estimating the potential revenue or demand at a new site is of the highest importance for making location decisions for businesses success in dynamic urban environments. Kriging is a popular method for estimating values at unvisited places, but calculations are only limited to the observed data in their neighbourhood. Hence a more comprehensive approach to formulating revenues is applying the underlying spatial interactions with their customers. Such spatial interaction modelling has a long history that primarily concerns the flows between origin and destination. While recently there have been some efforts to formulate spatial interaction models with the Bayesian framework they are only limited to flows at the disaggregated level [20, 25].

In Chapter 3, I developed a Bayesian spatial interaction model to simulate customers' behaviour with business facilities using their respective characteristics. BSIM considerably improves existing classical Huff type models as it formally addresses uncertainties arising in the modelling process via a Bayesian

framework while providing inferences at the level of business and customer locations. The key advantage of the proposed model is that it is scalable and can make inferences on large-scale datasets through variational inference, in contrast to the existing models. The BSIM is extended in Chapter 4 by lifting the assumption of fixed demand, which is also common in literature, by introducing dummy facilities to make more realistic estimations.

The synthetic experiments show how VI performs five times faster than MCMC while providing comparable performances in terms of parameter identification and without significant underestimation of the posterior covariance. It is important to consider all the data at a specific time horizon when fitting the model to assess the competition accurately. Therefore the usual cross-validation techniques are not applicable but can evaluate the model forecasts. In simulations studies, the model has proven to forecast future revenues by accounting for the changes in competition.

Addressing the vital question of best business facility location in competitive environments, I have formulated a mathematical modelling framework to simultaneously identify optimal facility locations and corresponding designs in a competitive environment in Chapter 4. This formulation considerably improves the existing competitive models based on classical utility methods as it considers model uncertainty via a Bayesian approach and provides probability density estimates of the revenue at new stores.

Additionally, I proposed a hierarchical search algorithm to overcome the challenge of providing exhaustive sets of potential locations to solve the optimisation problem in large geographical regions. The algorithm starts from an extensive collection of possible sites from a broad area and identifies the optimal facilities, then recursively explores the neighbouring locations until the objective value improvement is small. The first stage of the hierarchy can be executed in parallel to improve algorithm efficiency, but this could underrepresent the true combinations of optimal locations when searching for more than one facility. The initial candidate locations created with the multiresolution grid structure that accounts for density between customer spending and existing facilities reported the best and most efficient results.

One of the areas lacking in spatial interactions and facility location literature is that there are no real-world large scale applications. This is primarily because acquiring granular level real-world data is usually expensive. In Chapter 5, I presented datasets that include multiple variables observed at a granular level for public houses (pubs), supermarkets and customer zones at the postcode level. These datasets are formed by utilising both open and commercial proprietary data sources. The datasets are further enriched by adding reviews from Google's customer rating API, covering a broader audience than the traditional survey methods found in the literature.

For the first time, I have demonstrated the capability of estimating spatial interactions in large real-world urban areas such as Greater London with more than 1000 supermarkets, 1500 pubs and 150000 customer zones in Chapter 6. The inferences are made on different components of the spatial interactions, thereby making valuable conclusions for a business's ability to make decisions. As demonstrated in Chapter 1.2, clusters of customers can be inferred to identify customer segments that drive the sales at popular pubs. Hence the model can be of great value for businesses to understand their underlying revenue-generating mechanisms. Furthermore, BSIM is proven to outperform competing approaches in terms of prediction performances while providing consistent results with related indicators observed for the London region. However, the lack of time-series data limited the evaluation of the model forecasts in the real-world setting.

Subsequently, two case studies were presented to illustrate optimal sites: for a new company to enter the pubs market; and for the largest supermarket chain in the UK to expand its presence in the market. The optimal locations identified from the model demonstrate higher revenues than existing facilities while locating in less competitive areas, providing valuable insights for planning and decision-making. Although the introduced methodologies are presented only for supermarkets and pubs, they can also apply to any facility in the retail sector and other industries such as hospitality and healthcare. In the applications, we assume that the cost of locating is constant across the region, but considering spatial variation may produce more realistic results.

A leading prop-tech company in the UK, Nimbus property technology, supported the work in this thesis by providing valuable insights into the property industry and access to a more extensive database. The methodologies developed will be integrated into their system and make it available for property developers, business decision-makers and to a wider community to make more informed decisions on business locations in competitive urban environments.

In conclusion, my work stands out as the first effort to use a Bayesian framework to formulate revenue or demand for retail businesses based on their underlying generating spatial interactions with customers and thereby making location decisions. Additionally, the proposed variational inference advances existing models' capacity to deal with large-scale data. The methods provide valuable insights for planning and decision-making under uncertainty.

7.1.1 Code availability

BSIM is released under Apache License Version 2.0 and maintained in a public GitHub repository available at https://github.com/shanakap/BSIM. Detailed descriptions on how to run the codes are presented in the repository.

7.2 Limitations

A new approach to formulate customer-perceived utility of facilities is proposed in the context of spatial interaction models in Chapter 3. The probability density of isotropic Gaussian distribution is assumed to convey customers' utility by exploring the distance decaying nature in the distribution. Hence proximity between origin and destination to be of Euclidean distance, which is common in geostatistical models, although the urban landscape is unlikely to exhibit such properties. For example a store can be behind a customer location but entrance on different roads where the actual distance to travel is more than the Euclidean distance. Furthermore, the Gaussian distribution is assumed to isotropic where the variance of the distribution is same despite the direction. This may not hold in real world where the attraction levels could change across various directions.

One of the challenges in spatial modelling is the edge effects [66]. In this thesis, this is addressed in Chapter 3.2 by applying edge correction on the facilities' revenues to account for customers that are outside the study region. However, the competition created by the facilities outside the study area is not accounted for when edge corrections are applied. This is more evident from the results in chapter 6 where the optimal locations tend to be on the edge of the study area. Hence it is important to consider these limitations before deciding on the optimal sites presented in the study.

The formulation of the customers' budgeted spending (Eq. (3.5)) and the variance of the store-specific Gaussian distribution (Eq. (3.2)) assume a linear relationship with their respective features for simplicity. In practice, this assumption may not provide a realistic relationship; hence interpretation of the coefficients should be made with expert knowledge.

The BSIM formulates the spatial interactions that take place in a specific time horizon but does not account for the changes that influence over time. Hence the model has less information about the variation of revenues due to the changes that take place in the markets over time. This limits the use of the model for making accurate forecasts on facility revenue or demand.

In this thesis, the prior distributions are chosen to be weakly informative, and it remains to evaluate the impact of these priors. Hence, it is important to assess these priors' impact on the posterior through a sensitivity analysis. Therefore this limits the interpretations of the impact of prior distributions on the posterior distributions and predictions.

There are multiple Bayesian approaches proposed in this thesis where VI is proven to scale up well compared to MCMC technique. However, the computational time is also vital to generating real-time results in practice. The current setup cannot use mini-batch processing techniques since the model re-

quires all the data to correctly moderate the competition in the area. Hence this slows down computations and limits the ability to adapt techniques such as stochastic optimisation.

Furthermore, the VI framework for approximating the probability densities is focused on the mean-field inference (Eq. (3.11)) that assumes unknown variables are mutually independent. However, in practice, there may be dependencies between the approximate posterior distributions. Henceforth these assumptions limit the interpretation of the posterior distributions.

The introduced BSIM accounts for uncertainty and is expanded into making the location decisions to provide predictive distributions of the new sites in Chapter 4. However, in the optimisation process, the parameters' uncertainty is not accounted for but only uses the mean of the posterior distributions. Therefore the final results may not provide accurate predictions. Additionally, the cost of building facilities across greater London is assumed to be constant, but in practice, there are regions such as central London that are more expensive than the other areas. Henceforth this limits producing realistic outputs from the framework.

The spatial interaction model and the mathematical formulation of the competitive facility location introduced in this thesis are limited to learning one task at a time, for instance, focusing on the supermarket industry revenue only. Hence the model will not react to the changes in facilities from other sectors. Additionally, the optimal location framework is limited to identifying facilities from one industry at a time.

The modelling framework introduced in this thesis also encounters similar problems known in machine learning models. An important problem is the identification of model parameters as noticed in Chapter 3.3, and it is twofold [146] : (1) theoretical identification that arises from model specification usually associated with the presence of too many parameters; (2) empirical identification that results from lack of rich data for model estimation, although this is likely to improve with large data but not a sufficient condition [24]. The theoretical identification can lead to incorrect parameter estimation; hence important to explore ways to validate the parameter estimates from the model.

This thesis provides various datasets to overcome the empirical identification but has been restricted to modelling approximated revenues, and customers are represented by the postcode level. Thus the optimal locations provided in the real-world experiments may not provide accurate information. Furthermore, important features such as access to a road are not considered when evaluating the potential sites. Hence the results offer limited details to make the final decisions. Many of the limitations discussed can be overcome with the suggested future work in the next section.s

7.3 Further work

The proposed methodology can be extended and improved upon across multiple dimensions. First, one could consider adopting travel network to estimate the distance [30, 65, 87] instead of the Euclidean metric used in this study could provide a more realistic configuration of the geographical setting and lead to better inferences. Additionally, the model can be explored by applying other probability distributions such as Beta, Gamma and Wishart distribution to evaluate customers utility.

One of the important areas in spatial modelling is edge correction. There are avenues to improve the edge corrections introduced in chapter 3.2 by exploring methods to adjust for the competition due to the stores outside the study area. Furthermore, in this thesis, the extended BSIM is formulated in Chapter 4.2 to account for the unsatisfied demand in the market. Both the versions performed well with high R^2 (see Fig. 6.11 and Fig. 6.5), but further work is required to understand the practical implications and the inference on dummy facilities.

A better approximation of the budgeted spending of customers and variance of the distribution concerning their features is to adopt Gaussian Processes that would offer a more flexible framework but potentially less interpretable [122, 149]. This would provide better estimations for variance at new stores by excluding the unobserved parameter (ε_s).

The proposed framework can be extended to a spatio-temporal by considering the time in addition to the two geographic dimensions [36]. This approach could capture the time evolution of parameters to understand the behavioural changes of customers, and changes in urban systems will also be of significant interest. Furthermore, such advancements could provide better recommendations to place the new facilities by accounting for the changes in the urban systems.

The experiments should be expanded to improve the prior distributions. It may be beneficial to adopt priors derived from the sample data with methods such as maximum likelihood [31] or sample statistics [147]. Additionally, domain knowledge can be built into priors to restrict the range of a certain parameter or relationships between variables [144].

Further work is required to improve the computation time and explore the optimisation methods used in VI to assess the viability of modifying the algorithm to apply min-batch optimisation [94] techniques. Additionally, the current setup requires the model to be fitted to the entire dataset, but in practice, it will be beneficial to explore efficient ways to update the model parameters as new data becomes available.

In the VI framework, it adopts the mean-field to keep the computations

less complicated.The work can be expanded into more complex families to add dependencies between the variables, called structured variational inference [131]. This method potentially improves the approximations, but there is a trade-off as it is more difficult to solve the variational optimisation problem [18].

The proposed competitive facility location framework applies only the mean of the posterior distributions as for the parameters, but this could be extended to deal with uncertainty in the data of the optimisation problem by applying robust optimisation [13]. Furthermore, considering industry-led cost functions for placement or risk exposure is an interesting extension of my work that could be studied under Bayesian decision theory [12].

Further work is needed to exploit the advantages of multi-task learning that can solve multiple tasks simultaneously while learning commonalities and discrepancies across tasks [23]. In practice, this would be beneficial to learn from different industries and make optimal location predictions for multiple sectors simultaneously.

The real-world experiments can be improved by providing granular and accurate data. In practice, it is advantageous to explore the model performance with actual revenue or demand data from stores to gain rich information. Also, incorporating much granular level customer characteristics such as Experian Mosaic data [48] could improve richer data and provide more interpretable results. The store characteristics can be improved with features such as traffic data and available products at the stores. This thesis discusses two unique real-world case studies: a new company entering the market and a chain expanding existence in a competitive market. It remains to explore the formulation for an optimal facility location in a monopolistic market (Eq. (4.17)), such as the government expanding facilities with exclusive rights to provide certain services.

Adopting these directions in future studies can improve the BSIM to make accurate predictions while providing important inferences for decision making in businesses and the property industry.

Bibliography

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Robert Aboolian, Oded Berman, and Dmitry Krass. Competitive facility location model with concave demand. *European Journal of Operational Research*, 181(2):598–619, 2007.

[3] Luc Anselin. Spatial externalities, spatial multipliers, and spatial econometrics. *International regional science review*, 26(2):153–166, 2003.

[4] Roberto Basile, María Durbán, Román Mínguez, Jose María Montero, and Jesús Mur. Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. *Journal of Economic Dynamics and Control*, 48:229–245, 2014.

[5] David F Batten and David E Boyce. Spatial interaction, transportation, and interregional commodity flow models. In *Handbook of regional and urban economics*, volume 1, pages 357–406. Elsevier, 1987.

[6] Michael Batty. The size, scale, and shape of cities. *science*, 319(5864): 769–771, 2008.

[7] Michael Batty and S Mackie. The calibration of gravity, entropy, and related models of spatial interaction. *Environment and Planning A*, 4 (2):205–233, 1972.

[8] RD Bekti, N Pratiwi, and MT Jatipaningrum. Multiplicative competition interaction model to obtained retail consumer choice based on spatial analysis. In *IOP Conference Series: Earth and Environmental Science*, volume 187. IOP Publishing, 2018.

[9] Stefano Benati. The maximum capture problem with heterogeneous customers. *Computers & operations research*, 26(14):1351–1367, 1999.

[10] Stefano Benati and Pierre Hansen. The maximum capture problem with random utilities: Problem formulation and algorithms. *European Journal of Operational Research*, 143(3):518–530, 2002.

[11] RJ Bennett and RP Haining. Spatial structure and spatial interaction: Modelling approaches to the statistical analysis of geographical data. *Journal of the Royal Statistical Society: Series A (General)*, 148(1):1–27, 1985.

[12] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[13] James O Berger, Elías Moreno, Luis Raul Pericchi, M Jesús Bayarri, José M Bernardo, Juan A Cano, Julián De la Horra, Jacinto Martín, David Ríos-Insúa, Bruno Betrò, et al. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.

[14] Oded Berman and Dmitry Krass. Locating multiple competitive facilities: spatial interaction models with variable expenditures. *Annals of Operations Research*, 111(1-4):197–225, 2002.

[15] Oded Berman, Tammy Drezner, Zvi Drezner, and Dmitry Krass. Modeling competitive facility location problems: New approaches and results. In *Decision Technologies and Applications*, pages 156–181. INFORMS, 2009.

[16] Mark Birkin, Graham Clarke, and Martin Clarke. Refining and operationalizing entropy-maximizing models for business applications. *Geographical Analysis*, 42(4):422–445, 2010.

[17] C.M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer, 2006. ISBN 9780387310732.

[18] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[19] Kenneth E Boulding. General systems theory—the skeleton of science. *Management science*, 2(3):197–208, 1956.

[20] Matthew J Brierley, Jonathan J Forster, John W McDonald, and Peter WF Smith. Bayesian estimation of migration flows. *International Migration in Europe. Chichester, United Kingdom: Wiley*, pages 149–174, 2008.

[21] Stephen P Brooks and Andrew Gelman. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

[22] Gerald AP Carrothers. An historical review of the gravity and potential concepts of human interaction. *Journal of the American Institute of Planners*, 22(2):94–102, 1956.

[23] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[24] Elisabetta Cherchi and Juan de Dios Ortúzar. Empirical identification in the mixed logit model: analysing the effect of data richness. *Networks and Spatial Economics*, 8(2):109–124, 2008.

[25] Peter Congdon. Random-effects models for migration attractivity and retentivity: a Bayesian methodology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173(4):755–774, 2010.

[26] Cplex, IBM ILOG. User's manual for cplex 20.1.0, 2021. URL https://www.ibm.com/docs/en/icos/20.1.0?topic=cplex-users-manual.

[27] Noel Cressie. The origins of kriging. *Mathematical Geology*, 22(3):239–252, 1990. ISSN 08828121. doi: 10.1007/BF00889887.

[28] Noel Cressie. 4 - models for spatial processes. In John L. Stanford and Stephen B. Vardeman, editors, *Statistical Methods for Physical Science*, volume 28 of *Methods in Experimental Physics*, pages 93–124. Academic Press, 1994. doi: https://doi.org/10.1016/S0076-695X(08)60254-9.

[29] Noel Cressie and Gardar Johannesson. Fixed Rank Kriging for Very Large Spatial Data Sets. *Journal of the Royal Statistical Society, Series B*, 70(1):209–226, 2008. ISSN 1467-9868.

[30] Henry Crosby, Theo Damoulas, Alex Caton, Paul Davis, João Porto de Albuquerque, and Stephen A Jarvis. Road distance and travel time for an improved house price kriging predictor. *Geo-Spatial Information Science*, 21(3):185–194, 2018.

[31] William Francis Darnieder. *Bayesian methods for data-dependent priors*. PhD thesis, The Ohio State University, 2011.

[32] Luigi De Giovanni and Roberto Tadei. Modeling the retail system competition. *Procedia-Social and Behavioral Sciences*, 108:285–295, 2014.

[33] Rob Dekkers. *Applied systems theory*. Springer, 2015.

[34] Adam Dennett and Alan Wilson. A multilevel spatial interaction modelling framework for estimating interregional migration in europe. *Environment and Planning A*, 45(6):1491–1507, 2013.

[35] Department for Transport. National Public Transport Access Nodes (NaPTAN), 2014. URL https://data.gov.uk/dataset/ff93ffc1-6656-47d8-9155-85ea0b8f2251/national-public-transport-access-nodes-naptan.

[36] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72. IEEE, 2005.

[37] Tammy Drezner. Locating a single new facility among existing, unequally attractive facilities. *Journal of Regional Science*, 34(2):237–252, 1994.

[38] Tammy Drezner. Optimal continuous location of a retail facility, facility attractiveness, and market share: an interactive model. *Journal of retailing*, 70(1):49–64, 1994.

[39] Tammy Drezner. Derived attractiveness of shopping malls. *IMA Journal of Management Mathematics*, 17(4):349–358, 2006.

[40] Tammy Drezner. A review of competitive facility location in the plane. *Logistics Research*, 7(1):114, 2014.

[41] Tammy Drezner and Zvi Drezner. Modelling lost demand in competitive facility location. *Journal of the Operational Research Society*, 63(2):201–206, 2012.

[42] Tammy Drezner, Zvi Drezner, and P Kalczynski. Strategic competitive location: improving existing and establishing new facilities. *Journal of the Operational Research Society*, 63(12):1720–1730, 2012.

[43] Zvi Drezner, George O Wesolowsky, and Tammy Drezner. On the logit approach to competitive facility location. *Journal of Regional Science*, 38(2):313–327, 1998.

[44] Simon Duane, Anthony D Kennedy, Brian J Pendleton, and Duncan Roweth. Hybrid monte carlo. *Physics letters B*, 195(2):216–222, 1987.

[45] Pierre Dutilleul and Pierre Legendre. Spatial heterogeneity against heteroscedasticity: an ecological paradigm versus a statistical concept. *Oikos*, pages 152–171, 1993.

[46] Louis Ellam, Mark Girolami, Grigorios A Pavliotis, and A Wilson. Stochastic modelling of urban structure. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2213): 20170700, 2018.

[47] Glenn Ellison, Edward L Glaeser, and William R Kerr. What causes industry agglomeration? evidence from coagglomeration patterns. *American Economic Review*, 100(3):1195–1213, 2010.

[48] Experian. Experian mosaic data, May 2020. URL https://www.experian.co.uk/business/platforms/mosaic.

[49] José Fernández, Blas Pelegrı, Frank Plastria, Boglárka Tóth, et al. Solving a huff-like competitive location and design model for profit maximization in the plane. *European Journal of operational research*, 179(3): 1274–1287, 2007.

[50] José Fernández, Said Salhi, G Boglárka, et al. Location equilibria for a continuous competitive facility location problem under delivered pricing. *Computers & Operations Research*, 41:185–195, 2014.

[51] José Fernández, Juana L Redondo, Pilar M Ortigosa, G Boglárka, et al. Huff-like stackelberg location problems on the plane. In *Spatial Interaction Models*, pages 129–169. Springer, 2017.

[52] A Stewart Fotheringham. A new set of spatial-interaction models: the theory of competing destinations. *Environment and Planning A: Economy and Space*, 15(1):15–36, 1983.

[53] A Stewart Fotheringham and Michael J Webber. Spatial structure and the parameters of spatial interaction models. *Geographical Analysis*, 12 (1):33–46, 1980.

[54] Alexandre S Freire, Eduardo Moreno, and Wilfredo F Yushimito. A branch-and-bound algorithm for the maximum capture problem with random utilities. *European journal of operational research*, 252(1):204–212, 2016.

[55] Alan E Gelfand and Adrian FM Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409, 1990.

[56] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[57] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6:721–741, 1984.

[58] Charles J Geyer. Markov chain monte carlo maximum likelihood. 1991.

[59] Zoubin Ghahramani and Matthew J Beal. Propagation algorithms for variational Bayesian learning. *Advances in neural information processing systems*, pages 507–513, 2001.

[60] Francisco J Goerlich Gisbert. Weighted samples, kernel density estimators and convergence. *Empirical Economics*, 28(2):335–351, 2003.

[61] Edward Glaeser and José Scheinkman. Measuring social interactions. *Social dynamics*, pages 83–132, 2001.

[62] Oscar Gonzalez-Benito, Pablo A Munoz-Gallego, and Praveen K Kopalle. Asymmetric competition in retail store formats: Evaluating inter-and intra-format spatial effects. *Journal of Retailing*, 81(1):59–73, 2005.

[63] Google. Place Search, 2020. URL https://developers.google.com/places/web-service/search.

[64] Paul Gregory and Robert Stuart. *The global economy and its economic systems*. Nelson Education, 2013.

[65] Alexander Grigoryan. *Heat kernel and analysis on manifolds*, volume 47. American Mathematical Soc., 2009.

[66] Peter Haase. Spatial pattern analysis in ecology based on ripley's k-function: Introduction and methods of edge correction. *Journal of vegetation science*, 6(4):575–582, 1995.

[67] Pierre Hansen and Nenad Mladenović. Variable neighborhood search for the p-median. *Location Science*, 5(4):207–226, 1997.

[68] Hotelling Harold. Stability in competition. *Economic Journal*, 39(153): 41–57, 1929.

[69] Britton Harris and Alan Wilson. Equilibrium values and dynamics of attractiveness terms in production-constrained spatial-interaction models. *Environment and planning A*, 10(4):371–388, 1978.

[70] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

[71] Historic England. Listing, 2014. URL https://historicengland.org.uk/listing/the-list/.

[72] HM Land Registry. National polygon service, May 2020. URL https://www.gov.uk/guidance/national-polygon-service.

[73] HM Land Registry. Uk companies that own property in england and wales, 2020. URL https://www.gov.uk/guidance/hm-land-registry-uk-companies-that-own-property-in-england-and-wales.

[74] M John Hodgson. Toward more realistic allocation in location—allocation models: An interaction approach. *Environment and Planning A*, 10(11):1273–1285, 1978.

[75] M John Hodgson. A location—allocation model maximizing consumers' welfare. *Regional Studies*, 15(6):493–506, 1981.

[76] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

[77] Harold Hotelling. hotelling1990stability. *Economics Journal*, 39:41–57, 1929.

[78] David L. Huff. A Probabilistic Analysis of Shopping Center Trade Areas. *Land Economics*, 39(1):81, may 1963. ISSN 00237639. doi: 10.2307/3144521.

[79] David L Huff. A programmed solution for approximating an optimum retail location. *Land Economics*, 42(3):293–303, 1966.

[80] Rob J Hyndman and Anne B Koehler. Another look at measures of forecast accuracy. *International journal of forecasting*, 22(4):679–688, 2006.

[81] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[82] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[83] Hans Kellerer, Ulrich Pferschy, and David Pisinger. The multiple-choice knapsack problem. In *Knapsack Problems*, pages 317–347. Springer, 2004.

[84] Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? NIPS'17, page 5580–5590, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[85] Hande Küçükaydin, Necati Aras, and I Kuban Altınel. Competitive facility location problem with attractiveness adjustment of the follower: A bilevel programming model and its solution. *European Journal of Operational Research*, 208(3):206–220, 2011.

[86] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[87] John Lafferty, Guy Lebanon, and Tommi Jaakkola. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6(1), 2005.

[88] William Lazonick. *Business organization and the myth of the market economy*. Cambridge University Press, 1993.

[89] Giorgio Leonardi and Roberto Tadei. Random utility demand models and service location. *Regional Science and Urban Economics*, 14(3): 399–431, 1984.

[90] James P LeSage and Manfred M Fischer. Spatial econometric methods for modeling origin-destination flows. In *Handbook of applied spatial analysis*, pages 409–433. Springer, 2008.

[91] James P LeSage and Esra Satici. A Bayesian spatial interaction model variant of the poisson pseudo-maximum likelihood estimator. In *Spatial Econometric Interaction Modelling*, pages 121–143. Springer, 2016.

[92] James P LeSage, Manfred M Fischer, and Thomas Scherngell. Knowledge spillovers across europe: Evidence from a poisson spatial interaction model with spatial effects. *Papers in Regional Science*, 86(3): 393–421, 2007.

[93] PA W Lewis and Gerald S Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval research logistics quarterly*, 26(3): 403–413, 1979.

[94] Mu Li, Tong Zhang, Yuqiang Chen, and Alexander J Smola. Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 661–670, 2014.

[95] Yingru Li and Lin Liu. Assessing the impact of retail location on store performance: A comparison of wal-mart and kmart stores in cincinnati. *Applied Geography*, 32(2):591–600, 2012.

[96] Dennis Victor Lindley. *Bayesian statistics: A review*. SIAM, 1972.

[97] Ivana Ljubić and Eduardo Moreno. Outer approximation and submodular cuts for maximum capture facility location problems with random utilities. *European Journal of Operational Research*, 266(1):46–56, 2018.

[98] David JC MacKay and David JC Mac Kay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[99] Alfred Marshall. *Principles of economics*. Cosimo, Inc., 2009.

[100] Ministry of Housing, Communities & Local Government. English indices of deprivation 2019, sep 2019. URL https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019.

[101] Ministry of Housing, Communities and Local Government. English town centres, February 2020. URL https://data.gov.uk/dataset/ed07b21f-0a33-49e2-9578-83ccbc6a20db/english-town-centres-2004.

[102] Franco Modigliani and Richard Brumberg. Utility analysis and the consumption function: An interpretation of cross-section data. *Franco Modigliani*, 1(1):388–436, 1954.

[103] Alan T Murray. Evolving location analytics for service coverage modeling. *Geographical Analysis*, 50(3):207–222, 2018.

[104] Donald E Myers. Matrix formulation of co-kriging. *Journal of the International Association for Mathematical Geology*, 14(3):249–257, 1982.

[105] Masao Nakanishi and Lee G Cooper. Parameter estimation for a multiplicative competitive interaction model—least squares approach. *Journal of marketing research*, 11(3):303–311, 1974.

[106] Office for National Statistics. Lower layer super output area (lsoa) boundaries, nov 2016. URL https://data.gov.uk/dataset/fa883558-22fb-4a1a-8529-cffdee47d500/lower-layer-super-output-area-lsoa-boundaries.

[107] Office for National Statistics. National statistics postcode lookup, feb 2019. URL https://geoportal.statistics.gov.uk/datasets/4f71f3e9806d4ff895996f832eb7aacf.

[108] Margaret A Oliver and Richard Webster. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3):313–332, 1990.

[109] Margaret A Oliver and Richard Webster. *Basic steps in geostatistics: the variogram and kriging*. Springer, 2015.

[110] ONS. Retail sales index internet sales, aug 2020. URL https://www.ons.gov.uk/businessindustryandtrade/retailindustry/datasets/retailsalesindexinternetsales.

[111] Harmen Oppewal and Belinda Holyoake. Bundling and retail agglomeration effects on shopping behavior. *Journal of Retailing and Consumer services*, 11(2):61–74, 2004.

[112] Harmen Oppewal, Harry JP Timmermans, and Jordan J Louviere. Modelling the effects of shopping centre size and store variety on consumer choice behaviour. *Environment and Planning A*, 29(6):1073–1090, 1997.

[113] Ordnance Survey. AddressBase Premium, 2019. URL https://www.ordnancesurvey.co.uk/business-government/products/addressbase-premium.

[114] Ordnance Survey. Os mastermap topography layer, 2020. URL https://www.ordnancesurvey.co.uk/business-government/products/mastermap-topography.

[115] Maarten Ottens, Maarten Franssen, Peter Kroes, and Ibo Van De Poel. Modelling infrastructures as socio-technical systems. *International journal of critical infrastructures*, 2(2-3):133–145, 2006.

[116] Shanaka Perera, Theo Damoulas, Paul Davis, and Stephen Jarvis. Modelling business rates in england with big spatial data. In *In Proceedings of SIGKDD '19: International Workshop on Urban Computing*. SIGKDD, 2019.

[117] Shanaka Perera, Virginia Aglietti, and Theodoros Damoulas. On the competitive facility location problem with an extended Bayesian spatial interaction model. *Journal of the Royal Statistical Society. Series C: Applied Statistics*, Submitted on Dec 2021.

[118] Shanaka Perera, Virginia Aglietti, and Theodoros Damoulas. A Bayesian spatial interaction model for estimating revenue and demand at business facilities. *Nature Computational Science*, Submitted on Jan 2022.

[119] Jonne Pohjankukka, Tapio Pahikkala, Paavo Nevalainen, and Jukka Heikkonen. Estimating the prediction performance of spatial models via spatial k-fold cross validation. *International Journal of Geographical Information Science*, 31(10):2001–2019, 2017. ISSN 13623087. doi: 10.1080/13658816.2017.1346255. URL https://doi.org/10.1080/13658816.2017.1346255.

[120] Public Health England. Local alcohol profiles for england, February 2021. URL https://fingertips.phe.org.uk/profile/local-alcohol-profiles/.

[121] Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 814–822. PMLR, 2014.

[122] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[123] James Raymer. The estimation of international migration flows: A general technique focused on the origin-destination association structure. *Environment and Planning A*, 39(4):985–995, 2007.

[124] James Raymer, Alberto Bonaguidi, and Alessandro Valentini. Describing and projecting the age and spatial structures of interregional migration in italy. *Population, Space and Place*, 12(5):371–388, 2006.

[125] William J Reilly et al. Methods for the study of retail relationships. 1929.

[126] William John Reilly. *The law of retail gravitation*. New York, W.J. Reilly, 1931.

[127] Christian Robert and George Casella. *Monte Carlo statistical methods*. Springer Science & Business Media, 2013.

[128] Richard E Rossi, Jennifer L Dungan, and Louisa R Beck. Kriging in the shadows: geostatistical interpolation for remote sensing. *Remote Sensing of Environment*, 49(1):32–40, 1994.

[129] John R Roy and Jean-Claude Thill. Spatial interaction modelling. *Papers in Regional Science*, 83(1):339–361, 2003.

[130] Hanan Samet. *Foundations of multidimensional and metric data structures*. Morgan Kaufmann, 2006.

[131] Lawrence K Saul and Michael I Jordan. Exploiting tractable substructures in intractable networks. *Advances in neural information processing systems*, pages 486–492, 1996.

[132] Daniel Serra and Charles ReVelle. Competitive location and pricing on networks. *Geographical analysis*, 31(1):109–129, 1999.

[133] Bernard W Silverman. Density estimation for statistics and data analysis. *Monographs on Statistics and Applied Probability, London: Chapman and Hall*, 26, 1986.

[134] Stan Development Team. RStan: the R interface to Stan, 2020. URL http://mc-stan.org/. R package version 2.21.2.

[135] JCH Stillwell. Interzonal migration: some historical tests of spatial-interaction models. *Environment and Planning A*, 10(10):1187–1200, 1978.

[136] Andrew M Stuart. Inverse problems: a Bayesian perspective. *Acta numerica*, 19:451–559, 2010.

[137] Christoph Teller and Thomas Reutterer. The evolving concept of retail attractiveness: what makes retail agglomerations attractive when customers shop at them? *Journal of Retailing and Consumer Services*, 15 (3):127–143, 2008.

[138] Jean-Claude Thill. Research on urban and regional systems: Contributions from gis&t, spatial analysis, and location modeling. *Innovations in Urban and Regional Systems*, pages 3–20, 2020.

[139] Waldo R Tobler. A computer movie simulating urban growth in the Detroit region. *Economic geography*, 46(sup1):234–240, 1970.

[140] Constantine Toregas, Ralph Swain, Charles ReVelle, and Lawrence Bergman. The location of emergency service facilities. *Operations research*, 19(6):1363–1373, 1971.

[141] UK Data Service . Uk 2011 census postcode headcounts and households, feb 2011. URL https://www.statistics.digitalresources.jisc.ac.uk/dataset/uk-2011-census-postcode-headcounts-and-households-including-deprivation-ranks-individual.

[142] Valuation Office Agency. Valuation of public houses 2017. Technical Report Oct, 2016. URL https://www.gov.uk/government/publications/valuation-of-public-houses.

[143] Valuation Office Agency. Business rates, 2019. URL https://www.gov.uk/introduction-to-business-rates.

[144] Rens van de Schoot, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G Tadesse, Marina Vannucci, Andrew Gelman, Duco Veen, Joukje Willemsen, et al. Bayesian statistics and modelling. *Nature Reviews Methods Primers*, 1(1):1–26, 2021.

[145] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

[146] Joan Leslie Walker. *Extended discrete choice models: integrated framework, flexible error structures, and latent variables*. PhD thesis, Massachusetts Institute of Technology, 2001.

[147] Larry Wasserman. Asymptotic inference for mixture models by using data-dependent priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(1):159–180, 2000.

[148] Michael Wegener. Operational urban models state of the art. *Journal of the American planning Association*, 60(1):17–29, 1994.

[149] Christopher KI Williams and Carl Edward Rasmussen. Gaussian processes for regression. 1996.

[150] Alan Wilson. A family of spatial interaction models, and associated developments. *Environment and Planning A*, 3(1):1–32, 1971.

[151] Alan Wilson. Entropy in urban and regional modelling: Retrospect and prospect. *Geographical Analysis*, 42(4):364–394, 2010.

[152] Alan Geoffrey Wilson. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of transport economics and policy*, pages 108–126, 1969.

[153] Laurence A Wolsey and George L Nemhauser. *Integer and combinatorial optimization*, volume 55. John Wiley & Sons, 1999.

[154] Godwin Yeboah, João Porto de Albuquerque, Rafael Troilo, Grant Tregonning, Shanaka Perera, Syed AK Ahmed, Motunrayo Ajisola, Ornob Alam, Navneet Aujla, Syed Iqbal Azam, et al. Analysis of openstreetmap data quality at different stages of a participatory mapping process: Evidence from slums in africa and asia. *ISPRS International Journal of Geo-Information*, 10(4):265, 2021.

[155] Andrew Zammit-Mangion. *FRK: Fixed Rank Kriging*, 2018. URL https://CRAN.R-project.org/package=FRK. R package version 0.2.2.